AgRISTARS

DC-L2-04264
JSC-17829

A Joint Program for Agriculture and Resources Inventory Surveys Through Aerospace Remote Sensing

Domestic Crops and Land Cover

August 1982

EVALUATION OF SMALL AREA CROP ESTIMATION TECHNIQUES USING LANDSAT- AND GROUND-DERIVED DATA

M. L. Amis, M. V. Martin, W. G. McGuire, and S. S. Shen

Lockheed Engineering and Management Services Company, Inc.

Lyndon B. Johnson Space Center
Houston, Texas 77058

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DC-L2-04264; JSC-17829		

4. Title and Subtitle	5. Report Date
Evaluation of Small Area Crop Estimation Techniques Using Landsat- and Ground-Derived Data	August 1982
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
M. L. Amis, M. V. Martin, W. G. McGuire, and S. S. Shen	LEMSCO-17597
	10. Work Unit No.

9. Performing Organization Name and Address	11. Contract or Grant No.
Lockheed Engineering and Management Services Company, Inc. 1830 NASA Road 1 Houston, Texas 77258	NAS 9-15800
	13. Type of Report and Period Covered
12. Sponsoring Agency Name and Address	Final Report, 1980-81
National Aeronautics and Space Administration Lyndon B. Johnson Space Center Houston, Texas 77058 Technical Monitor: R. Heydorn	14. Sponsoring Agency Code

15. Supplementary Notes

The Agriculture and Resources Inventory Surveys Through Aerospace Remote Sensing is a joint program of the U.S. Department of Agriculture, the National Aeronautics and Space Administration, the National Oceanic and Atmospheric Administration (U.S. Department of Commerce), the Agency for International Development (U.S. Department of State), and the U.S. Department of the Interior.

16. Abstract

Tnis document describes the studies completed in fiscal year 1981 in support of the clustering/classification and preprocessing activities of the Domestic Crops and Land Cover project of the Agriculture and Resources Inventory Surveys Through Aerospace Remote Sensing program. The theme throughout the study was the improvement of subanalysis district (usually county level) crop hectarage estimates, as reflected in the following three objectives: (1) to evaluate the current U.S. Department of Agriculture Statistical Reporting Service regression approach to crop area estimation as applied to the problem of obtaining subanalysis district estimates, (2) to develop and test alternative approaches to subanalysis district estimation, and (3) to develop and test preprocessing techniques for use in improving subanalysis district estimates.

17. Key Words (Suggested by Author(s))		18. Distribution Statement
AgRISTARS	proportion estimator	
clustering	regression estimator	
crop estimator	spectral signature	
pixel		

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price*
Unclassified	Unclassified	145	

*For sale by the National Technical Information Service, Springfield, Virginia 22161

DC-L2-04264
JSC-17829

EVALUATION OF SMALL AREA CROP ESTIMATION TECHNIQUES USING LANDSAT- AND GROUND-DERIVED DATA

Job Order 71-352

This report describes the activities of the Domestic Crops and Land Cover project of the AgRISTARS program.

PREPARED BY

M. L. Amis, M. V. Martin, W. G. McGuire, and S. S. Shen

APPROVED BY

R. K. Lennington, Supervisor
Pattern Recognition Section

T. C. Minter, Jr., Manager
Supporting Research Department

LOCKHEED ENGINEERING AND MANAGEMENT SERVICES COMPANY, INC.

Under Contract NAS 9-15800

For

Earth Resources Research Division

Space and Life Sciences Directorate

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
LYNDON B. JOHNSON SPACE CENTER
HOUSTON, TEXAS

August 1982

LEMSCO-17597

PREFACE

The Agriculture and Resources Inventory Surveys Through Aerospace Remote Sensing is a multiyear program of research, development, evaluation, and application of aerospace remote sensing for agricultural resources, which began in fiscal year 1980. This program is a cooperative effort of the U.S. Department of Agriculture, the National Aeronautics and Space Administration, the National Oceanic and Atmospheric Administration (U.S. Department of Commerce), the Agency for International Development (U.S. Department of State), and the U.S. Department of the Interior.

The work which is the subject of this document was performed by the Earth Resources Research Division, Space and Life Sciences Directorate, Lyndon B. Johnson Space Center, National Aeronautics and Space Administration and Lockheed Engineering and Management Services Company, Inc. The tasks performed by Lockheed Engineering and Management Services Company, Inc., were accomplished under Contract NAS 9-15800.

CONTENTS

TABLES

FIGURES

ABBREVIATIONS AND ACRONYMS

AgRISTARS	Agriculture and Resources Inventory Surveys Through Aerospace Remote Sensing
CRD	Crop Report District
DC/LC	Domestic Crops and Land Cover
ERIM	Environmental Research Institute of Michigan
FY	fiscal year
MSE	mean squared error
MSS	multispectral scanner
SRS	Statistical Reporting Service
SSE	sum of squared error
USDA	U.S. Department of Agriculture

1. INTRODUCTION

1.1 OBJECTIVES

A major objective of the Statistical Reporting Service (SRS) of the U.S. Department of Agriculture (USDA) is the generation, with measurable precision, of accurate area estimates for crops and other land cover types. The areas of interest are national, regional, state, and various substate areas such as crop reporting districts (CRD's), groups of counties, and individual counties; currently, regression estimation is the method used, with Landsat classification results as the auxiliary variable of the estimator, and ground-observed data or ground truth from SRS operational surveys as the primary variable of the estimator. The ground truth is obtained by interviewing farm operators located in randomly selected areas of land called SRS segments. The regression estimator is defined over an analysis district, which is an area (usually a group of contiguous counties) in which the Landsat acquisitions used for estimation are the same for every point in the area. The area is "large" in the sense that it contains a sufficient number of SRS segments to reliably calculate regression coefficients.

This report documents the work done during fiscal year (FY) 1981 in the clustering and/or classification and preprocessing activities of the Domestic Crops and Land Cover (DC/LC) project of the Agriculture and Resources Inventory Surveys Through Aerospace Remote Sensing (AgRISTARS) program. The objectives of the research undertaken were threefold:

1. To evaluate the current SRS regression approach to crop area estimation when the area of interest is a single county or a small group of counties called a subanalysis district.
2. To develop and test new approaches to subanalysis district estimation.
3. To develop and test preprocessing techniques for use in improving subanalysis district estimation.

1.2 DISCUSSION OF OBJECTIVES

A subanalysis district is a subarea (usually a county) of an analysis district in which there is an insufficient number of SRS segments to reliably calculate regression coefficients.

The regression estimator can produce unbiased estimates with measurable precision for analysis districts; however, when the estimator developed over an analysis district is applied to a subanalysis district, it can be biased. The intent of the evaluation proposed in the first objective was to examine biasness and the applicability of an SRS-formulated estimator of the variance. The study consisted of empirically estimating the bias and variance of the subanalysis district estimator using a repeated sampling method. Reliable estimates of bias were thought to be possible because of an abundance of ground truth in some subareas. The empirical estimate of the variance would be compared to the formula-derived estimate, and, if possible, an improved subanalysis district variance estimator would be suggested.

An alternative regression approach developed by Manual Cardenas (ref. 1) was evaluated. The Cardenas family of estimators (section 2.2) was derived particularly for the case of small area estimation. Under certain assumptions, expressions for bias and variance of the estimators had been derived. Another class of estimators, referred to as direct proportion estimators, were also studied. These estimators did not depend on classification, but they estimated the posterior probability of a pixel belonging to a crop class. It was hoped that this approach would reduce bias, as well as variance, at the county level.

The focus of the preprocessing task was to effect some preliminary assessment of various preprocessing algorithms, which were developed in other studies to remove or reduce the variations in multispectral data resulting from changes in spectral signatures caused by sun angle, atmospheric conditions (including the presence of aerosals and water vapor), and background reflectance.

1.3 DESCRIPTION OF THE DATA SET

The data set used was from a six-county area in South Dakota, which comprised approximately 40 percent of one Landsat scene and which was previously used by the USDA in a soil study. The original data set included data from 252 segments; each segment was 65 hectares (160 acres, or one-fourth square mile) in area and had been chosen independently from 10 soil strata. Ground-truth data for these segments and registered Landsat data for two dates, July 26 and August 25, 1979, were supplied by the USDA. In its estimation procedure, the USDA typically uses 259-hectare (1-square-mile) segments randomly selected from land use strata. Because some soil strata were oversampled, resampling of the segments was necessary to more closely satisfy the requirements of this study. After resampling, 200 segments were available for the data set. These segments contained nine crop types that had sufficient numbers of pure pixels to train the classifier. There was some doubt concerning the sufficiency of the South Dakota data for estimating bias and variance using repeated sampling methods (see section 5).

2. A BRIEF DERIVATION OF THE ESTIMATORS AND ASSUMPTIONS

2.1 EDITOR SUBANALYSIS DISTRICT REGRESSION ESTIMATOR

A subanalysis district regression estimator was proposed by Huddleston and Ray (ref. 2), and it is the one referred to throughout this document as the current county-level estimator. It is, essentially, an analysis district regression estimator applied to a subarea of that analysis district; that is, regression coefficients are estimated using samples from the analysis district, whereas the mean being estimated is from a subpopulation of the analysis district. If the subpopulation is a set C of c counties (a sub-analysis district) then the separate form of the regression estimate of the total hectarage for C is:

$$\hat{Y}_{REG,c} = \sum_{k=1}^{L_c} N_{k,c}\left[\bar{y}_k + \hat{b}_k\left(\bar{X}_{k,c} - \bar{x}_k\right)\right] \qquad (1)$$

where

$N_{k,c}$ = the total number of area-frame units (segments) in the k^{th} stratum for the set C of c counties

L_c = the total number of strata for the set C of c counties

$\bar{y}_k$ = the average hectarage per sample unit from the ground survey for the k^{th} stratum for the crop of interest

$$= \sum_{j=1}^{n_k} y_{kj}/n_k$$

$\hat{b}_k$ = the estimated regression coefficient for the k^{th} stratum when regressing ground-truth hectarage on classified pixels for the n_k sample units

$$= \frac{\sum_{j=1}^{n_k}\left(x_{kj} - \bar{x}_k\right)\left(y_{kj} - \bar{y}_k\right)}{\sum_{j=1}^{n_k}\left(x_{kj} - \bar{x}_k\right)^2}$$

$\bar{X}_{k,c}$ = the average number of pixels per area-frame unit for all units in the k^{th} stratum for the set C of c counties that have been classified into the crop of interest

$$= \sum_{j=1}^{N_{k,c}} x_{kj}/N_{k,c}$$

$\bar{x}_k$ = the average number of pixels per sample unit in the k^{th} stratum that have been classified into the crop of interest

$$= \sum_{j=1}^{n_k} x_{kj}/n_k$$

The estimated variance of $\hat{Y}_{REG,c}$ has been proposed to be

$$v\left(\hat{Y}_{REG,c}\right) = \sum_{k=1}^{L_c} N_{k,c}^2 \left(\frac{N_k - n_k}{n_k}\right) S_{k,y}^2 \left(\frac{n_k - 1}{n_k - 2}\right) \cdot \left(1 - r_k^2\right) \left[I(C) + \frac{1}{n_k} + \frac{\left(\bar{X}_{k,c} - \bar{x}_k\right)^2}{\sum_{j=1}^{n_k} \left(x_{kj} - \bar{x}_k\right)^2} \right] \qquad (2)$$

where

N_k = the total number of area-frame units in the k^{th} stratum

n_k = the number of sample units in the k^{th} stratum

$S_{k,y}^2$ = the sample variance for the reported hectarage for the k^{th} stratum

$$= \sum_{j=1}^{n_k} \frac{\left(y_{kj} - \bar{y}_k\right)^2}{n_k - 1}$$

r_k^2 = the sample coefficient of determination for the k^{th} stratum

$I(C) = 1$ if C is a subset of the regression domain

$= 0$ if C is the entire regression domain

When $I(C) = 1$, the above variance formula is derived by treating the part of C contained in the k^{th} stratum as a single (fictitious) segment in which the number of pixels classified as the crop of interest is $\bar{X}_{k,C}$. This is equivalent to assuming that there is no variation at all for the actual segments in C. If there is such variation, then it is believed that the variance formula overestimates the variability of the subanalysis district regression estimator. Comparing the empirical variances with those obtained from the variance formula appears to substantiate this belief. For all of the major crops and for almost all of the minor crops, the empirical estimate of variance tends to be much closer to the formula variance for $I(C) = 0$ than for $I(C) = 1$, with most of the empirically observed values of I(C) falling in the interval [0, .1]. These results are found in section 5.

2.2 THE CARDENAS FAMILY OF ESTIMATORS

One of the problems encountered in estimating crop hectarage in a subanalysis district is that there may be few or no sample segments with which to obtain unbiased estimates of the mean hectarage per segment in the subanalysis district. Consider, for example, the six-county South Dakota area, and let C_k denote one of the counties. If $\bar{Y}_{kh}$ is the population mean hectarage per segment of a crop in land-use stratum h and in C_k, then the total for county k would be

$$Y_k = \sum_{h \epsilon C_k} M_{kh} \bar{Y}_{kh} \qquad (3)$$

where

$\sum_{h \epsilon C_k}$ = denotes the summation over all strata in county k

M_{kh} = the total number of segments in the h^{th} stratum within county k

An unbiased estimate of the $\bar{Y}_{kh}$ may not be possible if few sample segments belong to C_k; however, the analysis district does presumably contain sufficient sample segments to estimate $\bar{Y}_h$, the population mean crop hectarage per segment in stratum h. Thus, if the assumption that $\bar{Y}_{kh} = \bar{Y}_h$ were made, the total for county k would be estimated by

$$\hat{Y}_k = \sum_{h \epsilon C_k} M_{kh} \bar{Y}^*_h \tag{4}$$

where

$\bar{Y}^*_h = \frac{1}{n_h} \sum_{i=1}^{N_h} t_{ih} \bar{Y}^*_{ih}$, an unbiased estimate of $\bar{Y}_h$

$\bar{Y}^*_{ih} = \sum_{j=1}^{t_{ih}} y_{ihj}/t_{ih}$, the sample mean per segment of the area in the h^{th} stratum within county i

t_{ih} = the number of segments in the sample of the h^{th} stratum within county i

n_h = the number of counties in the sample of the h^{th} stratum

N_h = the number of counties in the h^{th} stratum.

Recognizing that the above assumption is not satisfied in general, Cardenas, Craig, and Blanchard (ref. 1) defined a family of county-level estimators using the classified pixels in each county and stratum as the auxiliary data. The family of estimators (referred to herein as the Cardenas family of estimators) for the k^{th} county is given by

$$\hat{Y}_{Bk} = \sum_{h \epsilon C_k} M_{kh} \left[\bar{Y}^*_h + B_h \left(\bar{X}_{kh} - \bar{X}_h \right) \right] \tag{5}$$

where

$\bar{X}_{kh}$ = the mean number of pixels classified as the crop in question for the h^{th} stratum within county k

$\bar{X}_h$ = the mean number of pixels classified as the crop in question for the h^{th} stratum

If $\bar{X}_{kh}$ is greater (less) than $\bar{X}_h$, then the mean area estimate should be increased (decreased) by an amount proportional to this difference. It follows that B_h should be positive.

If classification is such that $y_{ihj} = Ax_{ihj}$, where A is some constant, then using $B_h = \bar{Y}^*_h/\bar{X}_h$ yields an unbiased estimator (referred to as the Cardenas ratio estimator), $\hat{Y}_{rk}$ of Y_k.

Using a method similar to least squares estimation, estimates

$$B_h = \frac{M_h \sum_{i=1}^{N_k} t_{ih}\left(\bar{X}_{ih} - \bar{X}_h\right)\bar{Y}^*_{ih}}{n_h \sum_{i=1}^{N_h} M_{ih}\left(\bar{X}_{ih} - \bar{X}_h\right)^2}$$

yields an unbiased estimator (referred to as the Cardenas regression estimator), $\hat{Y}_{sk}$ of Y_k when $y_{ihj} = a + b_h x_{ihj}$, where a and b_h are constants.

The variances for $\hat{Y}_{rk}$ and $\hat{Y}_{sk}$ were derived by Cardenas et al. (ref. 1). If the assumption is made that the within-county variance is equal for all counties, then unbiased estimates of the variances were also given by Cardenas et al.

2.3 THE CLASSY-BASED DIRECT PROPORTION ESTIMATORS

One of the objectives of this study is to develop improved county-level crop area estimators. This may be achieved by modeling the county-level probability distribution as if it came from a mixture of distributions.

The general mixture model is given by

$$f(x) = \sum_{i=1}^{m} \alpha_i p(x|i) \qquad (6)$$

where

$p(x|i)$ = the probability density for distribution i

α_i = the proportion of distribution i in the mixture

m = the number of distributions in the mixture

$f(x)$ = the mixture probability density for a spectral value x

Applying the CLASSY clustering algorithm (ref. 3) to the unlabeled county-level data, it is possible to estimate m, $p(x|i)$, and α_i for $i = 1,\cdots,m$. The problem which remains is how to associate a crop label with each of the distributions, $p(x|i)$. This distribution labeling problem is the subject of a significant amount of ongoing research. Lennington and Terrell (ref. 4) described a maximum likelihood estimator for the proportion of a given distribution composed of a specific class. Chittineni (ref. 5) presented this maximum likelihood result and a similar result based on a probability of correct labeling criterion. Heydorn, Lennington, and Myers (ref. 6) presented a least squares, or regression, approach to this same problem. In each of these approaches the model is

$$\pi_k = \sum_{i=1}^{m} \beta_{ik}\alpha_i + \varepsilon \qquad (7)$$

where

π_k = the proportion of crop type k in the county of interest

β_{ik} = a set of "fitting" coefficients

α_i = the mixture proportions described previously

ε = error

Heydorn, Lennington, and Myers (ref. 6) have pointed out that this approach may be considered a generalization of stratified proportion estimation. Chittineni (ref. 5) observed that if the β's are restricted to either 0 or 1 (true distribution labeling), then the maximization problem may be solved exactly for the case of two or three subcrop types using an exhaustive search strategy.

All of these techniques for estimating the β_{ik} coefficients require that a small subset of labeled pixels be available. One way to select this subset of labeled pixels is to choose pixels from only those segments within the county of interest. This technique may not be feasible if the number of segments in the county of interest is small. Therefore, it seems appropriate to choose pixels also from segments within the county and adjacent to the county.

Not all of the approaches to obtaining estimates of the β_{ik} were evaluated. The chosen candidates were the maximum likelihood approach and the least squares, or regression, approach, both of which will now be discussed in more detail.

2.3.1 MAXIMUM LIKELIHOOD APPROACH

Suppose that the CLASSY clustering algorithm is applied to approximate the multivariate mixture density of the data in the county of interest. This results in a set of multivariate normal densities, $p(x|i)$, $i = 1,\cdots,m$, and a set of prior probabilities, α_i, $i = 1,\cdots,m$. Now, suppose that there is a set of data points, x_j, $j = 1,\cdots,n$, and let the random variable θ be the class label which takes on values of $\ell = 1,\cdots,c$. The joint probability of observing data point x_j associated with label $\theta = \ell$ may then be formulated as follows.

$$P(x_j,\theta = \ell) = \sum_{i=1}^{m} \alpha_i P(x_j,\theta = \ell|i)$$

$$= \sum_{i=1}^{m} \alpha_i P(\theta = \ell|x_j,i)P(x_j|i) \qquad (8)$$

Assume that $P(\theta = \ell|x_{j,i}) = P(\theta = \ell|i) = \beta_{\ell i}$, which means that the labeled random variable θ is conditionally independent of the observation x_j; i.e., given that one is sampling from distribution i, no further information about the class label is conveyed by knowing x_j.

Under this assumption, the proportion of class ℓ may be estimated as

$$\pi_\ell = P(\theta = \ell) = \sum_{i=1}^{m} \alpha_i \beta_{\ell i} \tag{9}$$

and $\beta_{\ell i}$ may be interpreted as the proportion of distribution i that is composed of class ℓ.

Now, a maximum likelihood approach may be used to estimate $\beta_{\ell i}$, assuming that all α_i and $p(x_j|i)$ are given.

Given a random sample of N ($= \sum_{\ell=1}^{c} N_\ell$) labeled data points from the county of interest, the likelihood function is

$$L = \prod_{\ell=1}^{c} \prod_{j_\ell=1}^{N_\ell} P\left(x_{j_\ell}, \theta = \ell\right) \tag{10}$$

where x_{j_ℓ}, $j_\ell = 1,\cdots,N_\ell$ represents those data points labeled as coming from class ℓ.

Under this mixture model, the likelihood function L may be written

$$L = \prod_{\ell=1}^{c} \prod_{j_\ell=1}^{N_\ell} \sum_{i=1}^{m} \alpha_i \beta_{\ell i} P(x_{j_\ell}|i) \tag{11}$$

To maximize L subject to the constraints $\sum_{\ell=1}^{c} \beta_{\ell i} = 1$ for $i = 1,\cdots,m$ is equivalent to maximizing the following function

$$F = \sum_{\ell=1}^{c} \sum_{j_\ell=1}^{N_\ell} \log\left(\sum_{i=1}^{m} \alpha_i \beta_{\ell i} p(x_{j_\ell}|i)\right) - \sum_{i=1}^{m} \eta_i \left(\sum_{\ell=1}^{c} \beta_{\ell i} - 1\right) \tag{12}$$

where η_i, $i = 1,\cdots,m$, is the Lagrange multiplier.

Maximizing with respect to $\beta_{\ell i}$, a solution of $\frac{\partial F}{\partial \beta_{\ell i}} = 0$ is given by

$$\hat{\beta}_{\ell i} = \frac{S_{\ell i}}{\sum_{\ell=1}^{c} S_{\ell i}} \tag{13}$$

where

$$S_{\ell i} = \sum_{j_\ell=1}^{N_\ell} \frac{\alpha_i \hat{\beta}_{\ell i} p(x_{j_\ell} | i)}{\sum_{i=1}^{m} \alpha_i \hat{\beta}_{\ell i} p(x_{j_\ell} | i)} \tag{14}$$

Therefore, $\beta_{\ell i}$ can be approximated using a fixed-point iteration scheme beginning with $\beta_{\ell i} = \frac{1}{c}$, $\ell = 1,\cdots,c$, $i = 1,\cdots,m$. Once the solution of $\beta_{\ell i}$ is obtained, the proportion of class ℓ can be estimated as

$$\hat{\pi}_\ell = \sum_{i=1}^{m} \alpha_i \hat{\beta}_{\ell i} \tag{15}$$

2.3.2 LEAST SQUARES APPROACH

Suppose again that the CLASSY clustering algorithm has been applied to approximate the multivariate mixture density of the data in the county of interest. This results in a set of multivariate normal densities, $p(x|i)$, and prior probabilities α_i, $i = 1,\cdots,m$. The model considered in this case is a regression model where $\beta_{\ell i}$ are just constants to be estimated, viz,

$$P(\theta = \ell | x_j) = \sum_{i=1}^{m} \beta_{\ell i} p(i | x_j) + \varepsilon \tag{16}$$

where

$P(\theta = \ell|x_j)$ = the posterior probability that x_j belongs to crop type ℓ

$p(i|x_j)$ = the posterior probability that x_j belongs to distribution i

ε = error

Now, the standard least squares techniques may be used to estimate $\beta_{\ell i}$. The criterion function to be minimized is

$$K = \| P(\theta = \ell | x_j) - \sum_{i=1}^{m} \beta_{\ell i} P(i | x_j) \|_F \tag{17}$$

where

$\| \cdot \|_F = \sqrt{\int (\cdot)^2 dF}$, and F is the cumulative distribution of the mixture density.

To minimize K is equivalent to minimizing

$$K^2 = \| P(\theta = \ell | x_j) - \sum_{i=1}^{m} \beta_{\ell i} P(i | x_j) \|_F^2$$

$$= \int \left[P(\theta = \ell | x_j) - \sum_{i=1}^{m} \beta_{\ell i} P(i | x_j) \right]^2 \left[\sum_{i=1}^{m} \alpha_i P(x_j | i) \right] dx_j \tag{18}$$

Minimizing with respect to $\beta_{\ell i}$, the solution is

$$\beta_{\ell i} = \sum_{k=1}^{m} q_{ik} E[P(\theta = \ell | x_j) P(k | x_j)]$$

where q_{ik} is the ik^{th} element of the inverse of the matrix

$$H = E[P(i | x_j) P(k | x_j)].$$

Given a random sample of labeled data points and associated labels $(x_j, \theta = \ell_j)$, $j = 1, \cdots, n$, where $\ell_j \in \{1, \cdots, c\}$, $\beta_{\ell i}$ can be estimated by

$$\hat{\beta}_{\ell i} = \sum_{k=1}^{m} \hat{q}_{ik} \frac{1}{n} \sum_{j=1}^{n} \psi_{\ell}(\ell_j) P(k | x_j) \tag{19}$$

where

$$\psi_{\ell}(\ell_j) = \begin{cases} 1 \text{ if } \ell = \ell_j \\ 0 \text{ otherwise} \end{cases}$$

$\hat{q}_{ik}$ = the ik^{th} element of $\hat{H}^{-1}$ and the ik^{th} element of $\hat{H}$ is

$$\frac{1}{n} \sum_{j=1}^{n} P(i | x_j) P(k | x_j)$$

The proportion of class ℓ is then estimated by

$$\hat{\pi}_{\ell} = \sum_{i=1}^{m} \alpha_i \hat{\beta}_{\ell i} \tag{20}$$

3. THE PREPROCESSING ALGORITHMS

3.1 XSTAR: AN ALGORITHM TO CORRECT LANDSAT DATA FOR THE EFFECTS OF HAZE AND SUN ANGLE

The XSTAR preprocessing algorithm is based on the Environmental Research Institute of Michigan (ERIM) radiative transfer model for an atmosphere with no absorption.

Letting primes denote a desired standard condition, the optical thickness for each multispectral scanner (MSS) channel I is represented as follows:

$$\tau_I = \tau'_{RI} + \alpha_I \gamma' \tag{21}$$

where

τ'_{RI} = the Rayleigh optical thickness in channel I

$\alpha_I \gamma'$ = the aerosol optical thickness in each channel so that γ' is a scalar measuring the amount of haze in the atmosphere in a hypothetical spectral band for which $\alpha_I = 1$

α_I = a function of the channel, independent of atmospheric haze

For Landsat-2 data, channels 1 through 4,

$$\underline{\alpha} = \begin{bmatrix} 1.2680 \\ 1.0445 \\ .9142 \\ .7734 \end{bmatrix} \tag{22}$$

The values for α_I were calculated from the estimated Landsat in-band optical thickness for an atmosphere with a horizontal visual range of 23 kilometers (14.38 miles), which is relatively clear.

Similarly, for an observed condition, the optical thickness is

$$\tau_I = \tau_{RI} + \alpha_I(\gamma' + \gamma) \tag{23}$$

However, the Rayleigh optical thickness is independent of atmospheric haze; so,

$$\left.\begin{aligned} \tau_{RI} &= \tau'_{RI} \\ \tau_I &= \tau'_I + \alpha_I \gamma \end{aligned}\right\} \quad \text{and} \tag{24}$$

The change in optical thickness from the standardized condition to be observed is then measured by γ.

If X_I is the observed and X'_I is the standardized Landsat radiance value in channel I, and assuming that other variables in the radiative transfer equation are restricted so that only atmospheric optical thickness is significant (ref. 8), a correction equation is obtained:

$$X'_I = e^{\alpha_I \gamma} X_I + \left(1 - e^{\alpha_I \gamma}\right) X^*_I + P(\alpha_I \gamma) \tag{25}$$

In general, both X^*_I and $P(\alpha_I \lambda)$ are functions of scanner geometry, illumination and viewing geometry, optical thickness, and the background albedo of the standardized conditions.

Excluding the higher order terms, represented by the polynomial function $P(\alpha_I \gamma)$,

$$X'_I = e^{\alpha_I \gamma} X_I + \left(1 - e^{\alpha_I \gamma}\right) X^*_I$$

or

$$(X'_I - X^*_I) = e^{\alpha_I \gamma} (X_I - X^*_I) \tag{26}$$

Then $\underline{X}^*$ specifies a point, or an origin, in the signal space relative to which the remainder of the signal space expands or contracts according to the effect of each multiplicative factor. For a sun angle of 39°,

$$\underline{X}^* = \begin{bmatrix} 61.9 \\ 66.2 \\ 83.2 \\ 33.9 \end{bmatrix} \tag{27}$$

To apply the XSTAR preprocessing algorithm to Landsat data, γ, a measurement of the amount of correction required, must be found. It is assumed that Landsat data distributions lie in a two-dimensional hyperplane in four-dimensional data space and the hyperplane position shifts with atmospheric haze. The XSTAR algorithm uses the tasseled-cap yellowness direction Y* as a measure of the component of the shift, which is perpendicular to the usual orientation of the plane. For the standardized condition, the average signal value measure in the $\hat{Y}$ direction is

$$Y^* = -11.2082 \text{ Landsat counts} \tag{28}$$

A value for γ that will shift the mean signal value ($\overline{X}_I$) is calculated so that the mean corrected signal value in the $\hat{Y}$ direction will be Y*.

$$Y^* = \sum_{I=1}^{4} \left[e^{\alpha_I \gamma} \frac{\mu_0'}{\mu_0} \overline{X}_I + \left(1 - e^{\alpha_I \gamma}\right) X_I^* \right] \hat{Y}_I \tag{29}$$

with $\mu_0' = \cos 39°$ and μ_0 = the cosine of the sun angle at time of acquisition.

If $e^{\alpha_I \gamma}$ is expanded and third and higher order terms are ignored,

$$a = \sum_{I=1}^{4} \alpha_I^2 \left[\frac{\mu_0'}{\mu_0} \overline{X}_I - X_I^* \right] \hat{Y}_I$$

$$b = \sum_{I=1}^{4} \alpha_I \left[\frac{\mu_0'}{\mu_0} \overline{X}_I - X_I^* \right] \hat{Y}_I \tag{30}$$

$$c = \sum_{I=1}^{4} \left[\frac{\mu_0'}{\mu_0} \overline{X}_I \hat{Y}_I \right] - Y^*$$

Then

$$\gamma \approx \frac{-b}{a} \left[1 - \sqrt{1 - \frac{2ac}{b^2}} \right] \tag{31}$$

For extremely hazy conditions, $1 - 2ac/b^2$ may be negative and the square root is set to zero; i.e.,

$$\gamma = \frac{-b}{a}$$

3.2 ATCOR: AN ALGORITHM TO CORRECT LANDSAT DATA FOR THE EFFECTS OF HAZE, SUN ANGLE, AND BACKGROUND REFLECTANCE

The ATCOR algorithm is designed to simulate the effects of target reflectance P_I, sun angle θ, haze level τ_H, and average reflectance of adjacent areas $\overline{P}_I$ on the radiance of a target as measured by Landsat in a channel I, and to correct for them. ATCOR assumes that radiances measured by the sensor can be modeled by

$$L_I = A_I(\overline{P}_I, \theta_0, \tau_H)P_I + B_I(\overline{P}_I, \theta_0, \tau_H) \qquad (32)$$

where L_I is the response in band I and A_I and B_I are coefficients for channel I which depend on $\overline{P}_I$, θ_0, and τ_H.

An atmospheric model was developed for use with ATCOR; the VandeHust method was then used to compute, for a range of wavelengths, the radiances gathered by the MSS for a range of values for $\overline{P}_I$, θ_0, τ_H, and P_I. These values are represented by a table in ATCOR.

Generally, θ_0 is known, but $\overline{P}_I$ and τ_H are not. However, if τ_H is known, then $\overline{P}_I$ can be calculated from the table.

The ATCOR program estimates τ_H, computes $\overline{P}_I$, and interpolates using the tables for $A_I(\overline{P}_I, \theta_0, \tau_H)$ and $B_I(\overline{P}_I, \theta_0, \tau_H)$ to find the correction coefficients which can be used to make the desired corrections.

The atmospheric model consists of two homogeneous layers: a Rayleigh scattering molecular layer on top and a Mie scattering haze layer next to the Earth's surface. Most haze is present in this region. The method used to determine the level of haze present actually estimates the total effect of all aerosols in the atmosphere and does not distinguish between haze and cirrus clouds. However, because the model assumes that this contribution is from haze particles in the lower atmosphere, the correction is less than optimal. Water vapor and other gaseous absorption are neglected.

The ATCOR program assumes that it is possible to obtain an estimate for the actual reflectance of the darkest pixels in a Landsat image and that the presence of haze will brighten the corresponding measurement at the sensor. The procedure for obtaining this estimate is discussed in reference 7. The atmospheric model indicates that the effect of haze is greatest in channel 1. The average channel 1 value for the darkest pixel in each scan line is computed (X_{min}). The reflectance of the darkest target is known or is set to a default value. From these values, the haze level which causes such a change between the actual or default (darkest) reflectance and the observed X_{min} is interpolated in the table. That value is the estimate for τ_H.

A_I and B_I may then be obtained from the table and the correction applied.

If primes denote the desired standard sun angle, haze level, and average background reflectance, then:

$$X_I' = A_I X_I + B_I \tag{33}$$

where

$$A_I = \frac{A_I(\overline{P}_I', \mu_0', \tau_H')}{A_I(\overline{P}_I, \mu_0, \tau_H)}$$

$$B_I = B_I(\overline{P}_I', \mu_0', \tau_H') - A_I B_I(\overline{P}_I, \mu_0, \tau_H)$$

$\mu_0 = \cos\ \theta$.

X_I' is the new radiance value for pixel X, channel I.

3.3 MLEST: A DISTRIBUTION MATCHING ALGORITHM

The MLEST algorithm is a statistical approach to finding an affine transformation of the form

$$Y = AX + B \tag{34}$$

which transforms clusters of normal distributions in the MSS signal space from a training area in a manner which best describes the clustering of distributions in a recognition area.

The objective of this approach is to model atmospheric and background effects using a maximum likelihood algorithm to develop a transformation matrix A and a vector B, in which the matrix A is not restricted to a diagonal matrix. This allows the estimated changes in a single MSS channel to be expressed as a weighted sum of the ensemble of channels rather than as a scalar transformation of only the data in that particular channel. This transformation is able to correct for haze differences and for any other affine transformations present in the data, regardless of origin. The primary advantages of MLEST over the XSTAR and ATCOR algorithms are that the nondiagonal terms of the transformation are included, and it is not necessary to make assumptions about minimum haze pixels.

The following procedure is used to evaluate the performance of the MLEST algorithm.

1. Obtain unlabeled clustering statistics for a training area. The overall probability density function for accomplishing this is

$$p(X_j) = \sum_{i=1}^{M} \alpha_i \, p(X_j|i) \qquad (35)$$

   where

   M is the number of clusters in the training area,

   X_j is the j^{th} pixel in the training area

   α_i is the proportion of the i^{th} distribution in the training area

2. Use these statistics and the MSS channel data from the recognition area as inputs to the MLEST algorithm. The MLEST algorithm estimates an affine transformation of the training statistics and the a priori cluster probabilities which maximize the likelihood function.

3. Transform labeled statistics from the training area using the computed affine transformation:

$$\hat{\underline{\mu}}_{i_R} = \hat{A}\underline{\mu}_{i_T} + \hat{\underline{B}}$$

$$\hat{\sum}_{i_R} = \hat{A}\sum_{i_T}\hat{A}' \qquad (36)$$

where

the subscripts R and T refer to the recognition and training areas, respectively.

$\hat{A}$ is the estimated transformation matrix

$\hat{\underline{B}}$ is the estimated transformation vector

$\underline{\mu}_i$ is the mean vector for the i^{th} distribution

$\sum_i$ is the covariance matrix for the i^{th} distribution

4. Use the transformed and labeled statistics to classify and label the pixels in the recognition area.

4. EXPERIMENT DESIGN DESCRIPTION

4.1 INTRODUCTION

The approach to be used was to estimate empirically the bias and variance of the estimator by repeated sampling. In order to implement this approach, it was necessary to determine the appropriate number of segments from the analysis district needed for training the classifier, that is, to determine the size of a training group. In addition, the number of training groups to be used had to be determined, and, if possible, these training groups made pairwise disjoint. Each training group would be used in the following ways:

1. A classifier would be trained using the segments in the training group.
2. The segments in the training group would be classified.
3. The regression coefficients for an estimator would be estimated using the ground-truth hectares and the number of classified pixels in the training group segments.
4. A given subanalysis district would be classified, and an estimate would be obtained of crop area in the subanalysis district, using the regression estimator in number 3.

The estimates of crop area obtained from the training groups would be used to calculate a sample estimate of variance and a mean estimate of crop area. The sample estimate of variance would be compared to the formula-obtained variance; and, as a measure of bias, the mean hectarage estimate would be compared to the direct expansion estimate, based only on the ground-truth segment data from the subanalysis district being estimated.

There was some question about the sufficiency of the South Dakota data for estimating bias and variance using the repeated sampling method just described. For comparison purposes, such a procedure should use repeated independent selections of segments for training; that is, the training groups should not overlap. A preliminary test study explored the issue of requiring training groups large enough for classification accuracy while at the same

time needing nonoverlapping training groups for the empirical estimation of bias and variance and for the use of subsequent statistical tests. This study is described in section 4.2.

4.2 FORMULATION OF GROUPS FOR TRAINING AND TESTING

Given the requirements imposed on the training groups by the repeated sampling method, a preliminary study was made to determine the appropriate size of the training groups for reliably estimating the mean and variance of the estimator. Some problems were anticipated and are now described.

The 252 65-hectare (one-fourth-square-mile) segments were obtained by sampling within-soil strata instead of land-use strata. Resampling, which was necessary because some strata were oversampled, reduced the number of available segments to 200. Ideally, a large number of independently selected and nonoverlapping groups of segments should be used with repeated sampling to do the empirical estimation. Because classification was also carried out in this study, each nonoverlapping group had to contain a sufficiently large number of segments to train the classifier. If the number of available segments is fixed, the number of segments within each nonoverlapping group decreases as the number of groups increases. Thus, if there were enough nonoverlapping groups to do the empirical estimation, these groups might not contain enough segments to adequately train the classifier. On the other hand, if there were enough segments in each nonoverlapping group to do the training, there might not be enough groups to do the empirical estimation. Therefore, it was apparent that, in order to have enough segments in each group to obtain acceptable classification performance and enough groups to conduct the empirical estimation of the variance, the use of overlapping groups was unavoidable. The training groups were determined with these constraints in mind. The county-level estimators were based on the requirements that a simple random sample of segments would be chosen within each land-use stratum and that the CLASSY clustering algorithm assumes a simple random sample from the population; thus, each of the soil strata that was

oversampled was resampled so that the new sample size for each stratum would be proportional to the area in that particular stratum. (Simple random sampling of a population is nearly equivalent to stratified random sampling with proportional allocation.) After resampling, 200 segments were left in the six-county area. These 200 segments were used to train the classifier, which was to be used as a benchmark in evaluating any other classifiers obtained in the repeated sampling process.

Previous experience in the FY 1980 DC/LC project indicated that seventy-five 65-hectare (one-fourth-square-mile) segments probably contained a sufficient number of pixels to train a classifier. Thus, the 200 segments were randomly partitioned into 8 sets containing 25 segments each and were denoted S_i, $i = 1,\cdots,8$. The training groups were formed by combining three sets at a time so that the intersection of any two training groups would be at most one set of 25 segments, and each would be used in exactly three different training groups. The collection of training groups used in this study is as follows:

$$\{S_1 \cup S_2 \cup S_3,\ S_1 \cup S_4 \cup S_6,\ S_2 \cup S_4 \cup S_7,\ S_2 \cup S_5 \cup S_8,$$
$$S_3 \cup S_4 \cup S_5,\ S_3 \cup S_6 \cup S_8,\ S_5 \cup S_6 \cup S_7,\ S_1 \cup S_7 \cup S_8\}$$

Some of the advantages of combining the partitions to obtain training groups instead of using simple random sampling are:

1. The maximum number of overlapping segments in any two groups can be controlled.
2. Each segment is chosen the same number of times, whereas in simple random sampling some segments may never be chosen and some could be chosen more than the others.

Each of the training groups was used to train a classifier. Then the entire six-county area was classified and county-level estimates were obtained. The variability of the eight classifiers was examined, and the performances were compared with the benchmark of training on all 200 segments.

The criteria on which the collection of training groups was accepted were:

1. Individual classification performances did not depart significantly from the benchmark.
2. The number of groups was large enough to provide reliable empirical estimation.

4.3 QUESTIONS ADDRESSED IN THE EVALUATION STUDIES

The evaluation study for the current county-level estimator addressed two questions.

First, when the value I(C) = 1 is used in the variance formula, the resulting number was believed to be an overestimate of the variability of the current county-level estimator. Recall from section 2.1 that I(C) = 1 is used whenever C is a proper subset of the regression domain and is equivalent to assuming that there is no variation at all for the segments in C (the county). If C is the entire regression domain, then I(C) = 0; and the estimator is simply the current analysis district regression estimator. Obtaining an empirical estimate of the variance and comparing it to the formula variance using different values for I(C) would be a means of examining the above assumption and also of estimating a more realistic value for I(C).

The second question was whether or not the current county-level estimator was an unbiased estimator of the total crop hectarage for a county. To answer this question decisively would require knowing the true crop hectarage for a county, and this information was not available. Instead, the standard used for comparison was the direct expansion estimator for the county, $\hat{Y} = N \times \overline{y}$, where N is the total number of possible segments in the county and $\overline{y}$ is the sample mean crop hectarage per segment for the given county.

The particular alternative county-level estimators that were evaluated were of interest because of the approaches that were taken in hectarage estimation at the subanalysis district level. The Cardenas family of estimators compares the average number of crop pixels per segment in a given stratum to the

average number of crop pixels per segment in the given county in that stratum and adjusts the mean area estimate by an amount proportional to this difference. Therefore, it was desirable to compare the performance of two of the members of this family to the current county-level estimator as well as to compare the performances of these two Cardenas estimators. In addition, the only variance formula available for this family makes the assumption that, for all counties, the within-county variances are equal. To compare the empirical estimate of variance to the formula variance as an indication of the validity of this assumption was also a desirable objective.

The two direct proportion estimators offered the possibility of estimating crop hectarage in a county using only the county Landsat data and relatively few ground-truth pixels. Another advantage was that the classification of each pixel is not done directly as is necessary for the regression type estimators. For making comparisons, however, this was also a disadvantage in this study. For each county, one estimate of a crop proportion was obtained rather than eight estimates using eight classifier training groups. Questions regarding the size of bias and variance were answered by using the proportions and variances generated by the simple random sample approach (section 5.3) as the standard.

4.4 PREPROCESSING

The objective of the preprocessing study was to see if candidate preprocessing algorithms applied to analysis district Landsat imagery have the capability for improving crop area estimates at the county level when few (or no) training segments are available from that county. Three preprocessing algorithms were chosen for study based on results of the Large Area Crop Inventory Experiment: XSTAR, ATCOR, and MLEST (see section 3). The XSTAR and ATCOR algorithms are haze-correction models which transform the analysis district and the county to be estimated to correct for the presence of haze and/or background effects and to make them look spectrally similar to the classifier. The MLEST algorithm takes distributions present in the analysis district and estimates an affine shift correction which matches them to distributions from the county. A

transformation is obtained which may then be used on the statistics in the classifier before classifying the county.

Ideally, sample segment data chosen from the analysis district (a six-county area) would be used as the training set on which to develop the regression estimator, and an entire county would be used as the test area. However, ground truth was available only for sample segments in the county. In this study, we tested whether preprocessing improves the estimates for a sample from a county, rather than whether it improves an actual county estimate.

In order to address the worst possible case, two test areas (sample segments from Beadle and Kingsbury counties, South Dakota) that did not overlap the training set were chosen. This effort not to duplicate sample segments from the training set in the county was made for two reasons: First, to achieve distinct test and training groups for the F-test and the Hotelling T^2 test; and, second, to provide the "worst" case, where no sample segments from the area of interest were available for training. This selection also fulfilled the requirement that sample segments from surrounding areas be available for training; the surrounding segments in this case were other training group segments that were in Beadle and/or Kingsbury Counties.

After estimates for the county samples were obtained by the USDA EDITOR system and by the USDA EDITOR with MLEST, ATCOR, and XSTAR preprocessing, a comparison was made to see which method produced estimates closer to the true ground-truth proportions.

The purpose was to ascertain if one of the preprocessing methods had in some way made the regression estimator, which was developed over the analysis district, appropriate and accurate at the county level.

4.5 STATISTICAL EVALUATION APPROACH

It was apparent from the preliminary analysis that the overlap among some training groups would vitiate any statistical tests requiring that assumptions of independence of random variables be satisfied. This was accepted as a

necessary flaw in order to have enough training groups with enough segments in each group to adequately train the classifier.

In evaluating the conjecture that the formula for the variance of the current subanalysis district regression estimator overestimates the variability of this estimator if I(C) = 1 is used, the following approach was taken: The arithmetic means over the eight training groups were plotted against the corresponding values of I(C) = 0 and I(C) = 1 for each crop. The line containing these two points expresses the linear relationship existing between I(C) and the variance of $\hat{Y}_c$, as is evident from the formula for the variance of $\hat{Y}_c$. This line can be used to approximate the value of I(C) associated with the empirical estimate of variance.

The Behrens-Fisher test was used to investigate the bias of the current county-level estimator. Whenever a sample of segments is randomly selected from a county, the direct expansion estimator $N \times \overline{y}$ is an unbiased estimator for the total county hectarage of a given crop. Likewise, each individual y_i is unbiased for the mean number of hectares per segment. Similarly, if the current county-level estimator is unbiased for the total county hectarage of a given crop, then this estimator divided by the number of segments in the county (N) is unbiased for the mean number of hectares per segment for a given crop. The Behrens-Fisher test indicates whether the current county-level estimator, when divided by N, systematically overestimates or underestimates the mean number of hectares per segment.

For the Cardenas ratio and regression estimators, the proportions of each crop in each county, as well as a "coefficient of variation" for each crop, are presented in tables. These figures are calculated for each training group. This "coefficient of variation" is, for each county, the ratio of the square root of formula variance for the crop divided by the average number of hectares of the crop. In addition, a sample coefficient of variation was obtained using the estimates of a crop from the eight training groups as samples. These summary statistics for each estimator are presented by crop.

The sample variances of the Cardenas estimators were compared, and the sample variance of each Cardenas estimator was compared to the sample variance of the current county-level regression estimator using an F-test. This was done knowing that the independence assumptions were not satisfied. Indeed, not only did some of the training groups overlap, but also all three estimators have the same y-variable, namely the ground truth hectarage per segment. However, it was believed that a comparison of the sample variances would indicate whether or not they were significantly different.

The Behrens-Fisher t-test described previously was the test for bias of the Cardenas ratio and regression estimators.

The bias and the mean squared error (MSE) of the direct proportion estimators were calculated and recorded as summary statistics. Recall that the procedure is to cluster a county and obtain proportions, α_i, of each distribution in the mixture model. Then, 500 labeled pixels are chosen randomly from segments within the county to estimate β_{ℓ_i}, or the proportion of crop ℓ in distribution i. The proportion of crop ℓ, π_ℓ, is $\sum_{i=1}^{c} \alpha_i \beta_{\ell_i}$, where c is the number of distributions present. In addition, the crop proportions taken from the 500 pixels are computed to obtain a third estimate, called the simple random sample estimate. This procedure is repeated 50 times, each time choosing 500 labeled pixels randomly from the segments within the county. The average number of labeled pixels available in each county is about 6700, a large enough number that the 50 repetitions can be considered independent. The proportion of each crop, determined using all the labeled pixels in a county, is considered the true proportion of that crop. In order to estimate the bias, the 50 estimates are averaged and the mean compared to the proportion in the labeled pixels. The MSE is the sample variance of the 50 proportion estimates. An F-ratio is computed for each of the direct proportion estimators. This is the ratio of the sample variance of the direct proportion estimator to the sample variance of the simple random sample estimates over the 50 repetitions. Independence problems are again present, since each of the three estimates (maximum likelihood, least squares, and simple random sample estimates) is obtained over the same set of 500 pixels.

4.6 EVALUATION OF PREPROCESSORS

The comparison of the performance of each preprocessor with the USDA EDITOR was done using the Hotelling T^2 test, which compares the mean difference between ground truth and the regression estimate per segment for both methods. Accepting the null hypothesis would indicate that there is no significant difference between estimates produced by the two methods.

There remains some question as to whether the regression equation developed from the analysis district should be used on the county in evaluating the performance of the MLEST algorithm, since a new (transformed) classifier is used on that county. Although an improvement in classification was obtained using the MLEST algorithm on the county data, a corresponding improvement in estimation might not occur using the regression lines developed on the analysis district if these regression lines are not appropriate for the county. So in addition to the Hotelling T^2 test for the other preprocessors, two other tests were made: one to compare regression estimators for the USDA EDITOR and MLEST, which were developed on the county; and one to compare estimates from the USDA EDITOR using the training regression lines and MLEST using the county regression lines. This issue is discussed in further detail in section 5, Study Results.

If the results of the Hotelling T^2 test show that one or more of the preprocessing procedures produce estimates that are not significantly different from those produced by the USDA EDITOR alone, it is necessary to examine the mean vector to determine if the results of the preprocessing procedure are better, worse, or mixed. If the estimates using the preprocessor are closer to ground truth for every crop than those of the EDITOR alone, then the results using the preprocessing procedure are considered better; if they are further from ground truth for every crop than those of the EDITOR alone, the results using the preprocessing procedure are considered worse. If the estimates using the preprocessor are closer to ground truth for some crops and further for others, it may be concluded that one procedure is not better than the other.

In order to attempt to detect the presence or absence of haze or other differences between the test and training areas, a two-sided F-test for homogeneity of variances and an F-test for equality of analysis district and county regression lines is done for each crop. These tests are discussed in more detail in section 5.

5. STUDY RESULTS

5.1 CURRENT SUBANALYSIS DISTRICT REGRESSION ESTIMATOR

5.1.1 EXPLANATION OF GRAPHS AND TABLES

Figures 5-1 through 5-9 contain plots of variance versus I(C) for the current county regression estimator. For each crop, the formula variance using I(C) = 1 is computed for each training group, and an average is obtained. Similarly, the formula variance using I(C) = 0 is computed for each training group, and an average is obtained. These two numbers determine the line associated with each crop. The empirical estimate of variance is then used with this linear relationship to produce an empirical estimate of I(C).

Although these plots have been produced for only one county, other data exhibited later provide a similar result: for the majority of crops in each county, the empirically estimated values of I(C) are around zero. This tends to confirm the statement that the formula variance provides an estimate which greatly overestimates the variance of the current county-level estimator.

5.1.2 THEORETICAL AND EMPIRICAL VARIANCE ESTIMATES BY COUNTY

Tables 5-1 through 5-6 present the preceding graphical results quantitatively by county. The averages across training groups of the theoretical and empirical variance estimates for each crop are given. The empirically observed value of I(C) is also given, and it was determined by observing that in the formula I(C) is linearly related to the variance estimate. The averages of the variance estimates with I(C) = 1 and also with I(C) = 0 provide two points determining the line representing this linear relationship. By using this fact and the empirical estimate of variance, one can obtain a corresponding value of I(C). For the majority of crops in each county, these values of I(C) are close to zero.

TABLE 5-1.- THEORETICAL AND EMPIRICAL VARIANCE ESTIMATES
USING CURRENT USDA PROCEDURE FOR BEADLE COUNTY

[Hectares × 10^6]

Item	Rangeland	Sunflowers	Corn	Wheat	Oats	Grass	Alfalfa	Hay cut	Flax
Average of variance estimates with I(C)=1	2428.5	160.0	580.4	1034.5	532.3	969.5	680.3	322.6	285.7
Empirical estimate of variance	332.9	1.1	4.2	131.9	25.9	60.0	76.7	36.8	5.2
Average of variance estimates with I(C)=0	79.9	4.5	19.5	32.5	15.8	25.6	21.7	14.4	9.1
Empirically observed value of I(C)	.11	-.02	-.03	.10	.02	.04	.08	.07	-.01

TABLE 5-2.- THEORETICAL AND EMPIRICAL VARIANCE ESTIMATES
USING CURRENT USDA PROCEDURE FOR CLARK COUNTY

[Hectares × 10^6]

Item	Rangeland	Sunflowers	Corn	Wheat	Oats	Grass	Alfalfa	Hay cut	Flax
Average of variance estimates with I(C)=1	1160.9	75.5	293.3	496.2	261.7	414.2	335.3	186.3	135.6
Empirical estimate of variance	41.3	3.2	5.0	5.5	16.0	46.3	20.6	16.4	28.2
Average of variance estimates with I(C)=0	40.0	2.4	10.6	15.1	10.0	11.3	12.8	10.7	5.7
Empirically observed value of I(C)	.001	.01	-.02	-.02	.02	.09	.02	.03	.17

TABLE 5-3.- THEORETICAL AND EMPIRICAL VARIANCE ESTIMATES
USING CURRENT USDA PROCEDURE FOR CODINGTON COUNTY

[Hectares × 10^6]

Item	Rangeland	Sunflowers	Corn	Wheat	Oats	Grass	Alfalfa	Hay cut	Flax
Average of variance estimates with I(C)=1	504.5	32.5	130.5	216.7	116.9	166.0	148.6	87.9	57.9
Empirical estimate of variance	48.1	.7	5.1	5.4	31.8	26.3	7.0	5.5	22.3
Average of variance estimates with I(C)=0	20.5	1.1	5.1	7.0	6.7	4.7	6.3	5.9	3.0
Empirically observed value of I(C)	.06	-.01	0.0	-.008	.23	.13	.005	-.006	.35

TABLE 5-4.- THEORETICAL AND EMPIRICAL VARIANCE ESTIMATES USING CURRENT USDA PROCEDURE FOR HAMLIN COUNTY

[Hectares × 10^6]

Item	Rangeland	Sunflowers	Corn	Wheat	Oats	Grass	Alfalfa	Hay cut	Flax
Average of variance estimates with I(C)=1	404.6	30.7	107.3	187.7	94.9	117.8	131.1	69.9	40.6
Empirical estimate of variance	18.9	0.1	7.5	3.2	29.2	7.9	7.1	2.3	11.8
Average of variance estimates with I(C)=0	17.5	1.1	5.6	6.4	3.6	3.5	5.5	2.5	2.1
Empirically observed value of I(C)	.004	-.03	.02	-.02	.28	.04	.01	-.003	.25

TABLE 5-5.- THEORETICAL AND EMPIRICAL VARIANCE ESTIMATES
USING CURRENT USDA PROCEDURE FOR KINGSBURY COUNTY

[Hectares × 10^6]

Item	Rangeland	Sunflowers	Corn	Wheat	Oats	Grass	Alfalfa	Hay cut	Flax
Average of variance estimates with I(C)=1	1522.4	100.8	349.3	638.5	327.7	663.1	412.2	175.4	187.2
Empirical estimate of variance	50.0	0.6	13.8	15.6	10.1	17.0	19.7	2.6	77.0
Average of variance estimates with I(C)=0	44.0	2.7	11.0	17.1	9.1	16.8	11.2	4.8	7.4
Empirically observed value of I(C)	.004	-.02	.01	-.002	.003	0.0	.02	-.01	.39

TABLE 5-6.- THEORETICAL AND EMPIRICAL VARIANCE ESTIMATES
USING CURRENT USDA PROCEDURE FOR SPINK COUNTY

[Hectares × 10^6]

Item	Rangeland	Sunflowers	Corn	Wheat	Oats	Grass	Alfalfa	Hay cut	Flax
Average of variance estimates with I(C)=1	3058.8	213.1	756.9	1364.9	693.2	1109.8	923.8	421.5	328.9
Empirical estimate of variance	209.4	65.2	17.3	115.7	64.1	35.5	276.4	61.6	24.2
Average of variance estimates with I(C)=0	89.6	11.8	24.8	42.8	21.6	31.2	36.2	22.3	9.9
Empirically observed value of I(C)	.04	.27	-.010	.05	.06	.004	.27	.1	.04

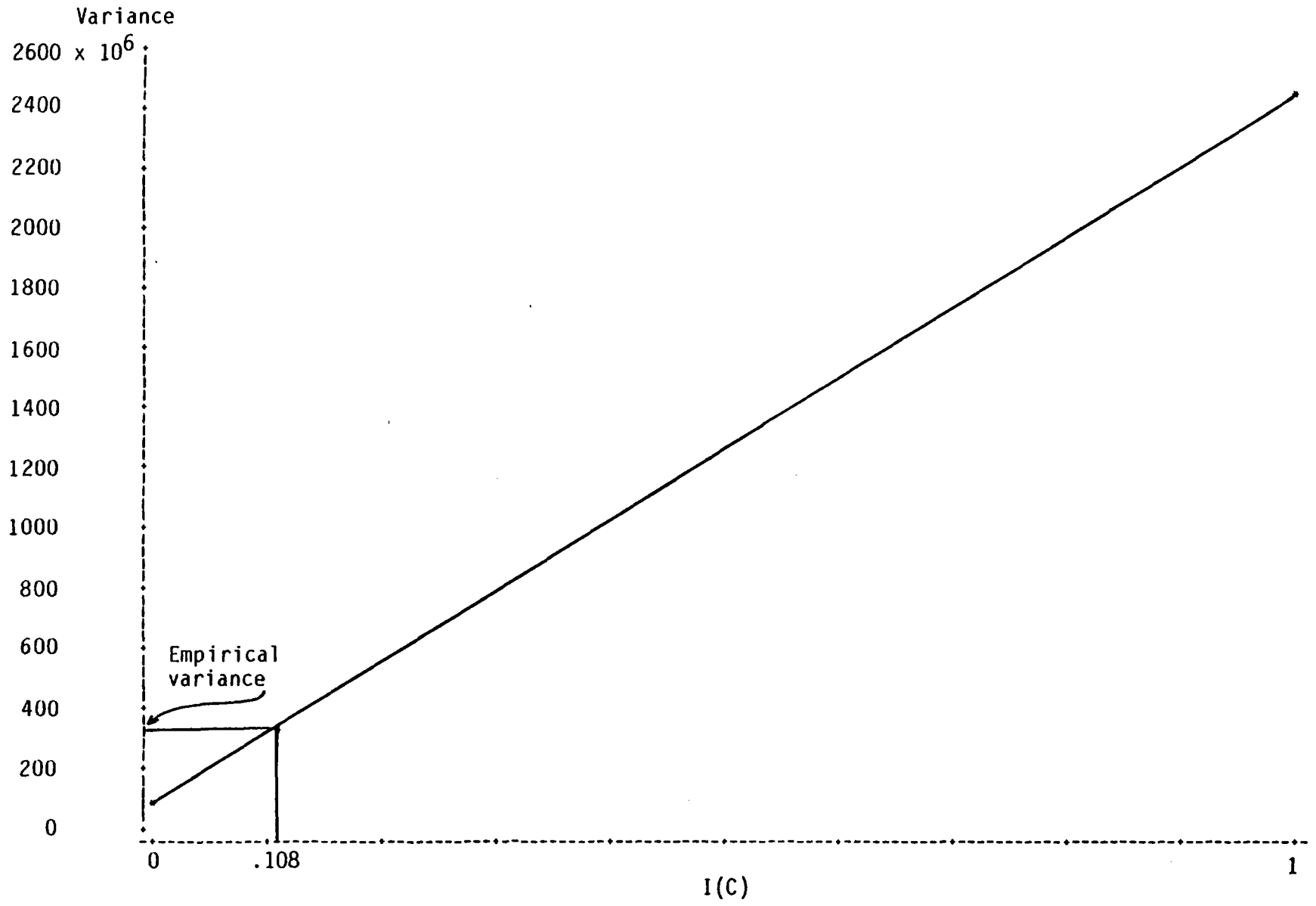

Figure 5-1.- Variance versus I(C) for rangeland in Beadle County.

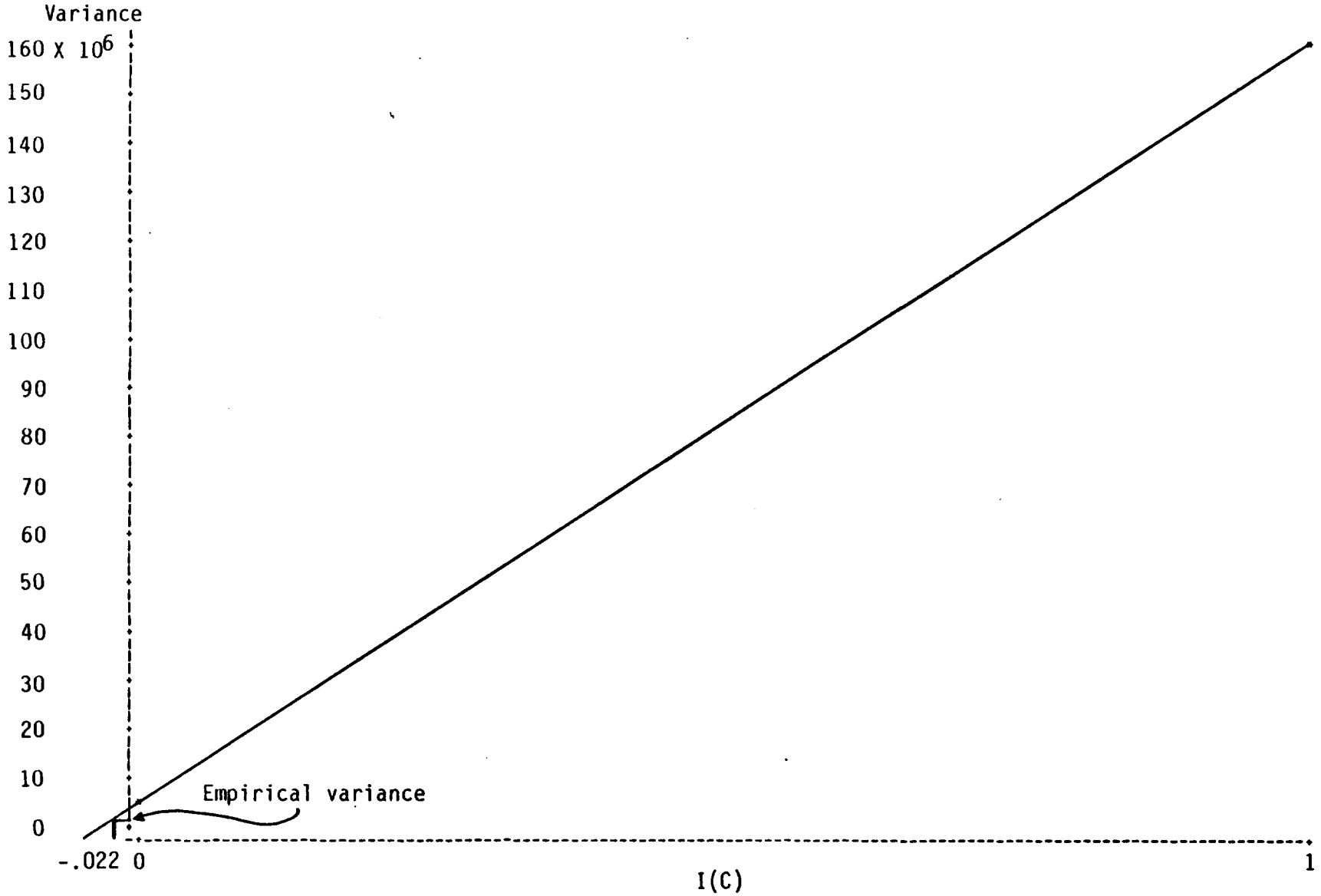

Figure 5-2.- Variance versus I(C) for sunflowers in Beadle County.

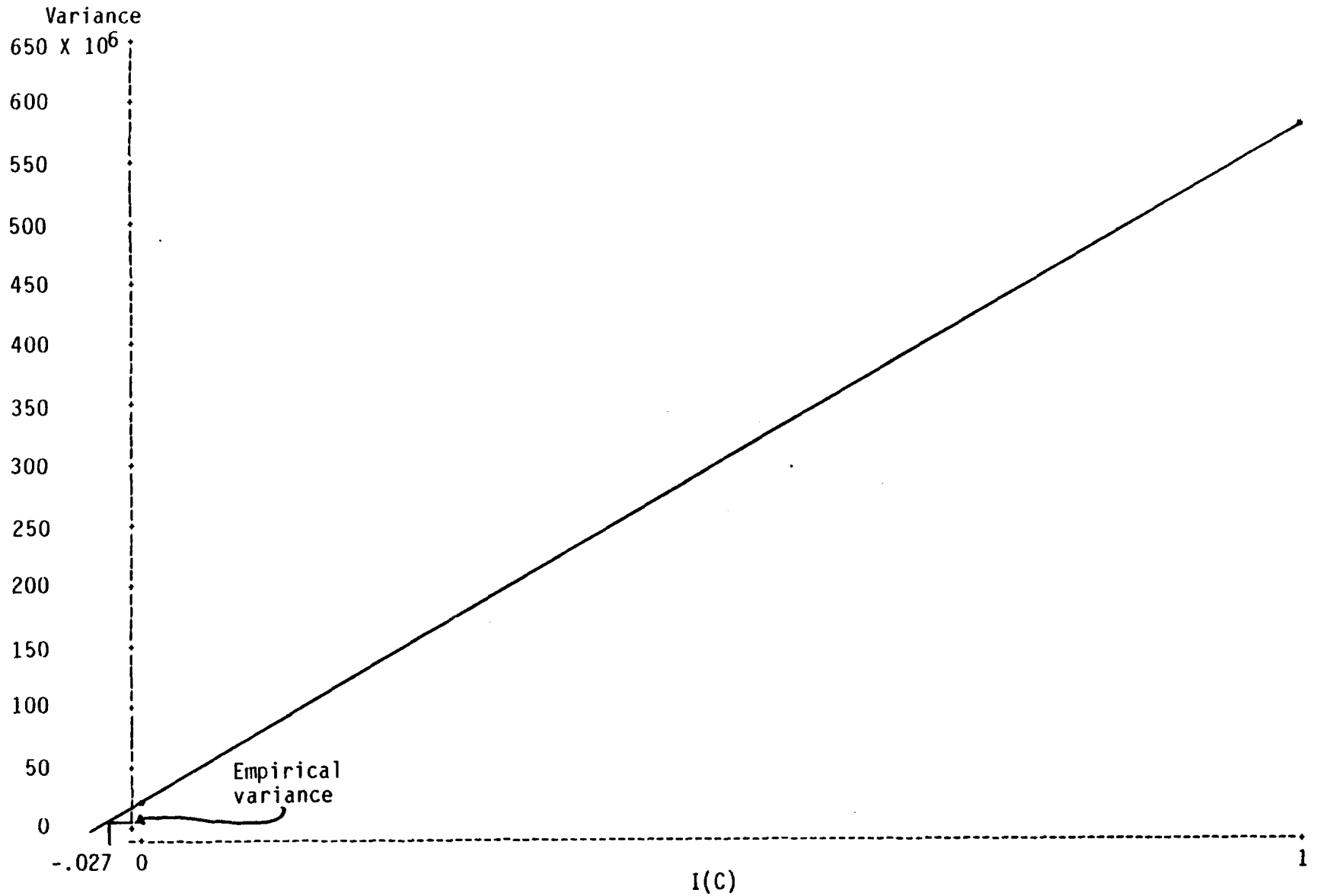

Figure 5-3.- Variance versus I(C) for corn in Beadle County.

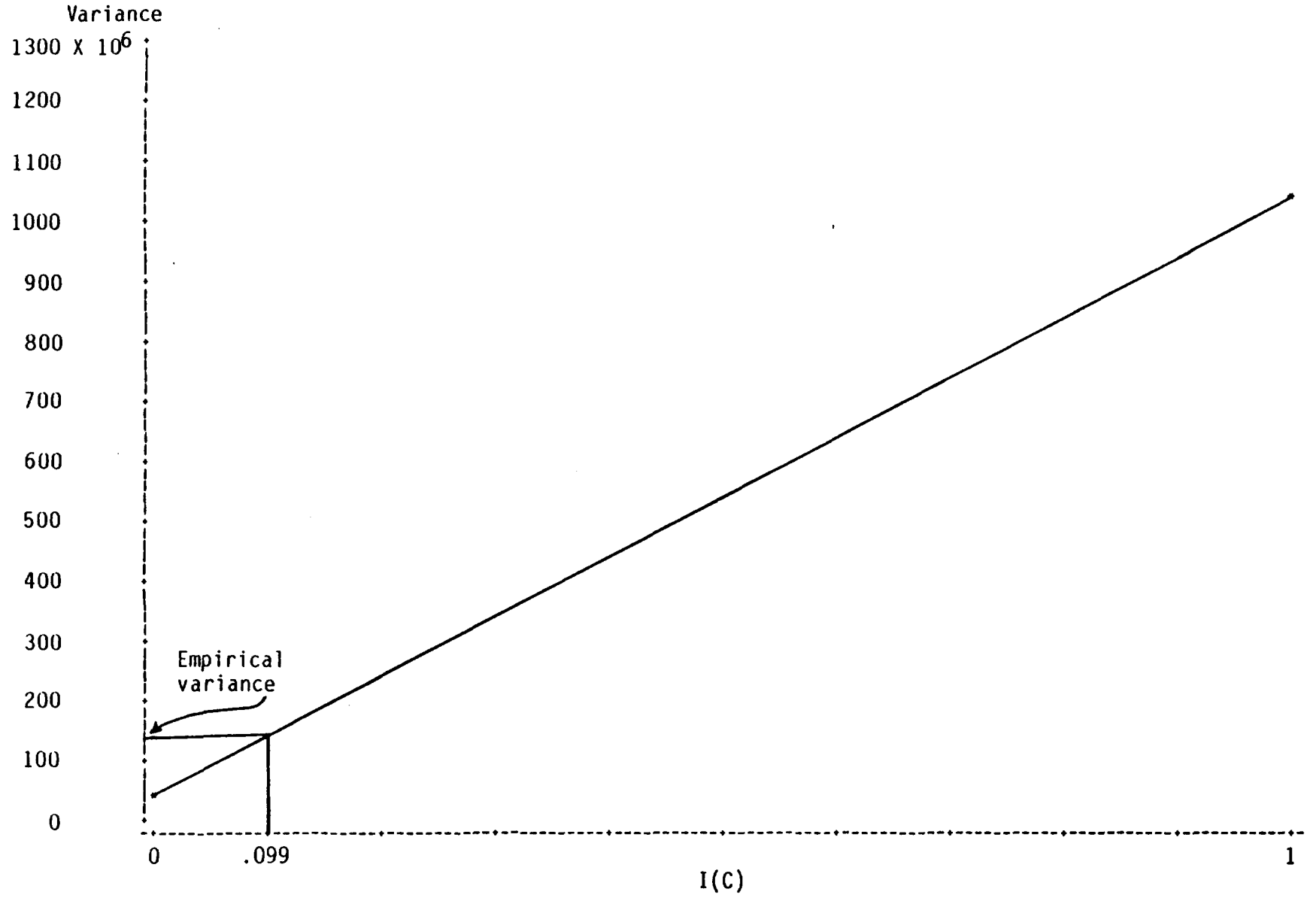

Figure 5-4.- Variance versus I(C) for wheat in Beadle County.

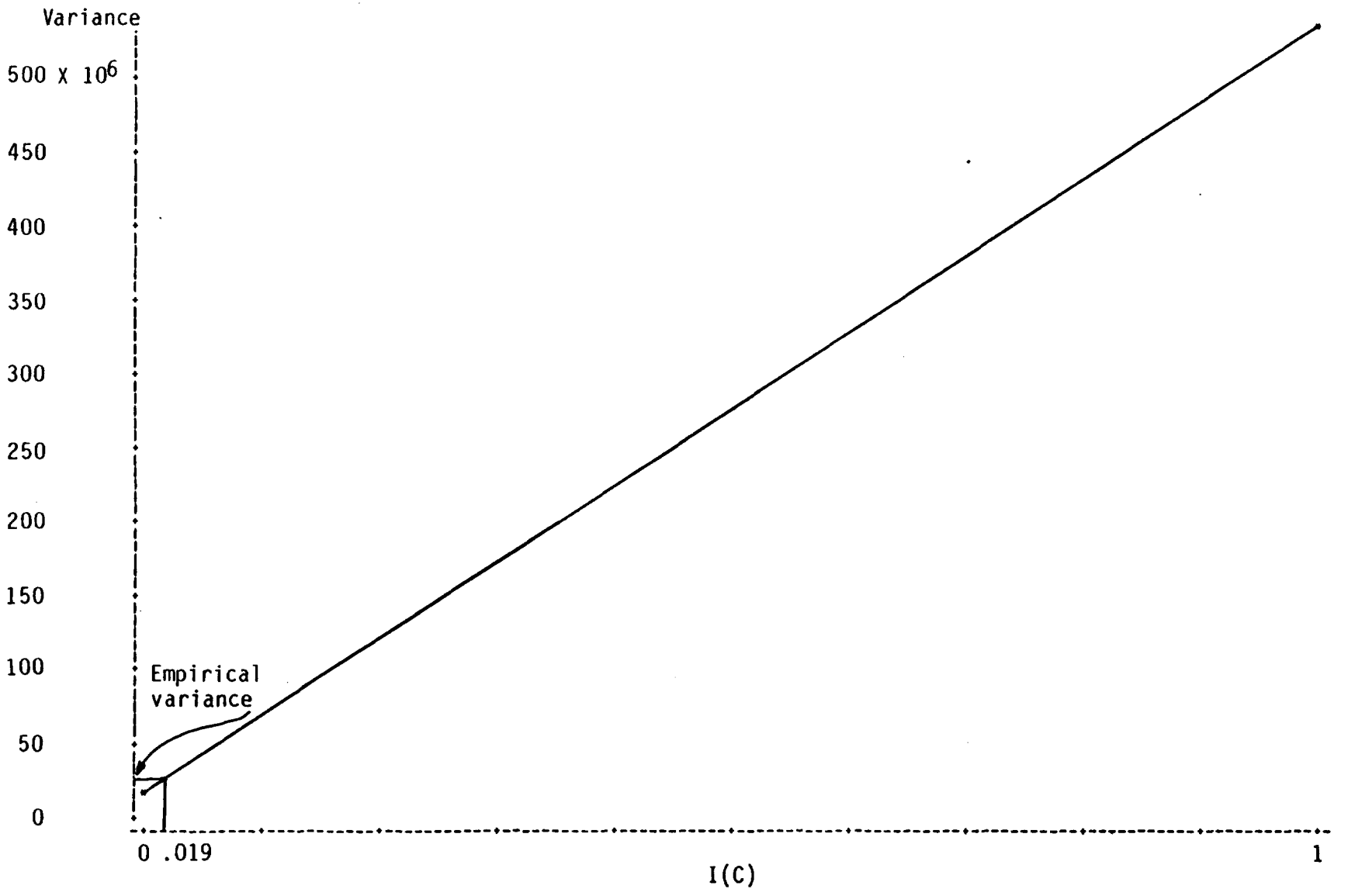

Figure 5-5.- Variance versus I(C) for oats in Beadle County.

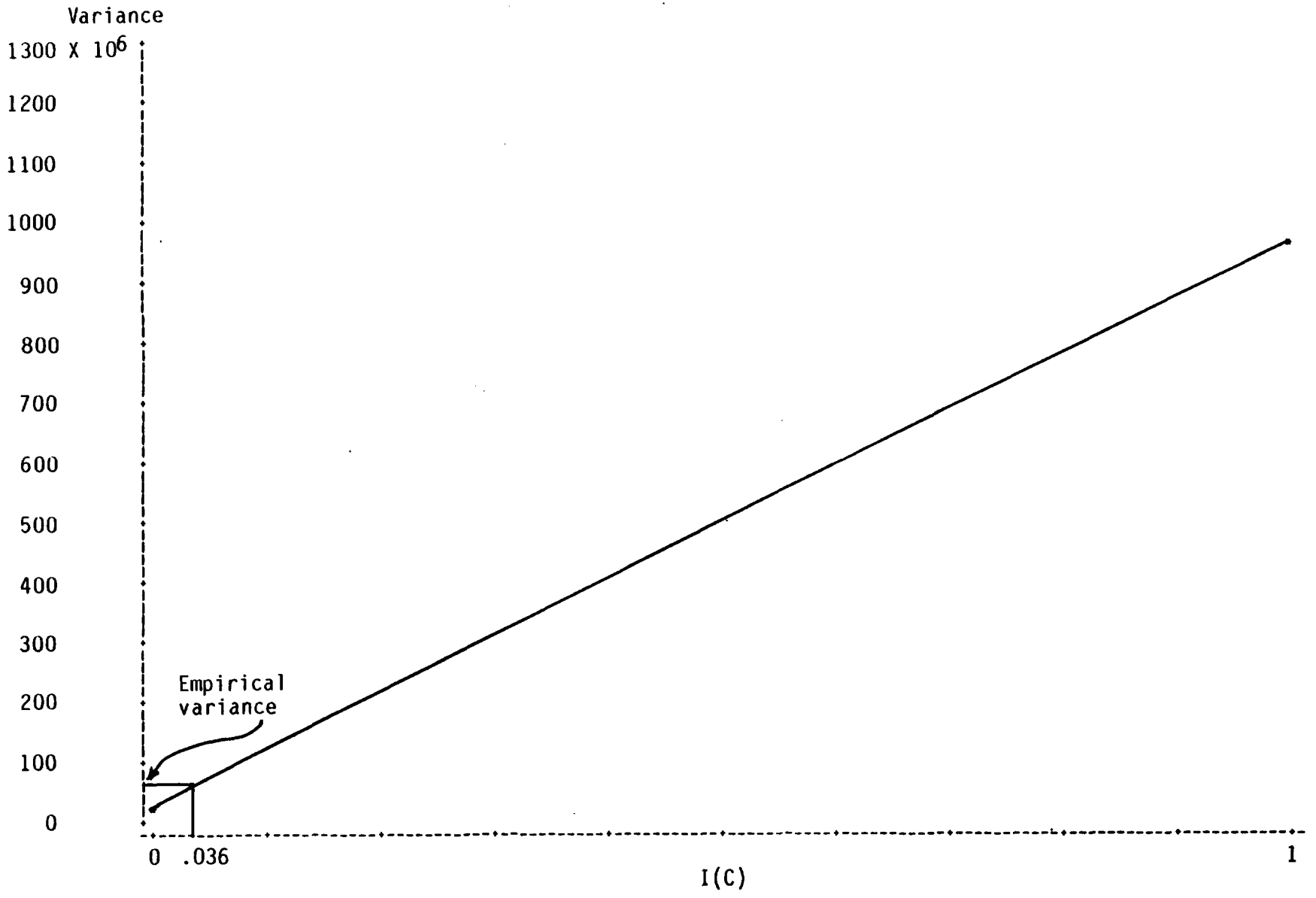

Figure 5-6.- Variance versus I(C) for grass in Beadle County.

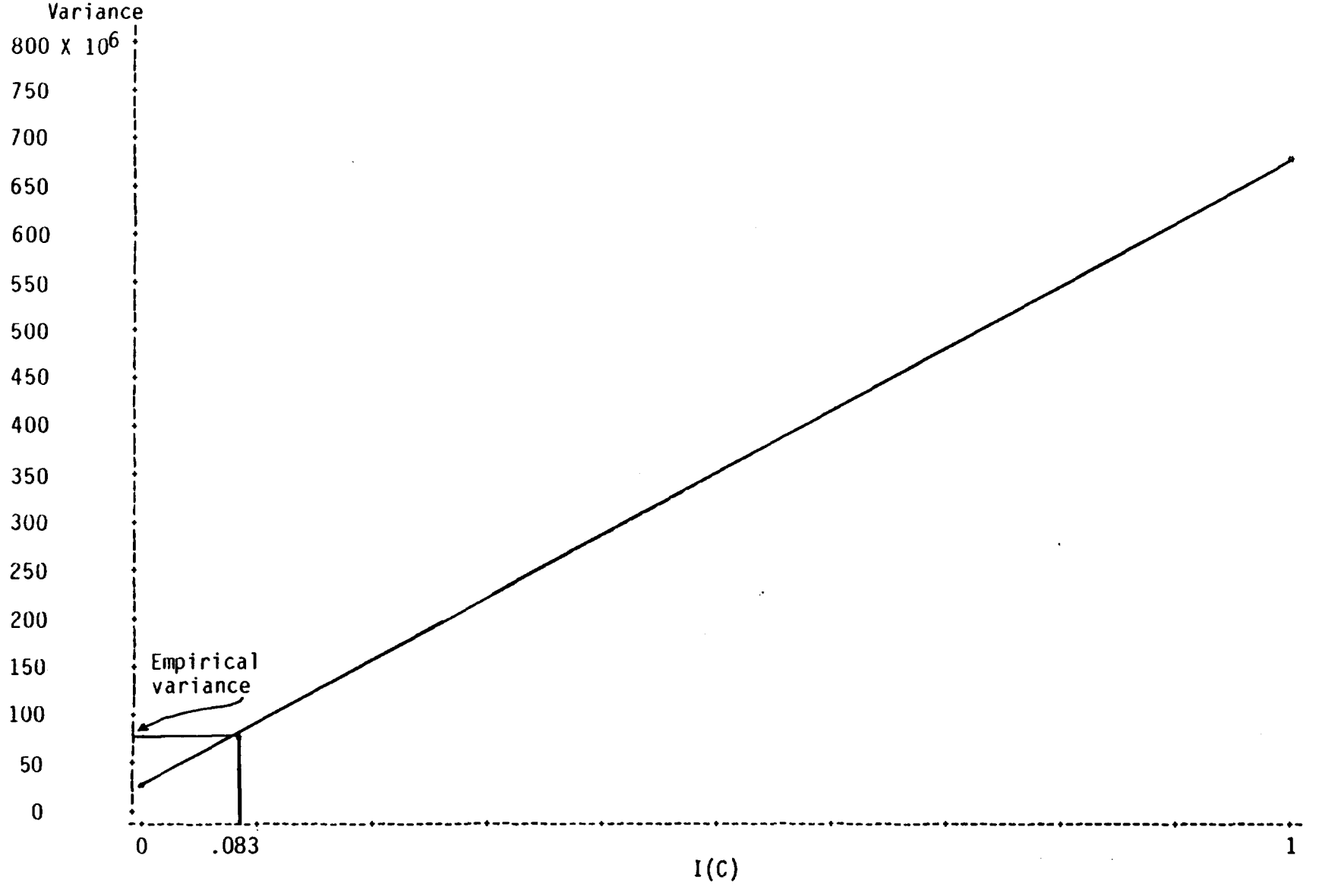

Figure 5-7.- Variance versus I(C) for alfalfa in Beadle County.

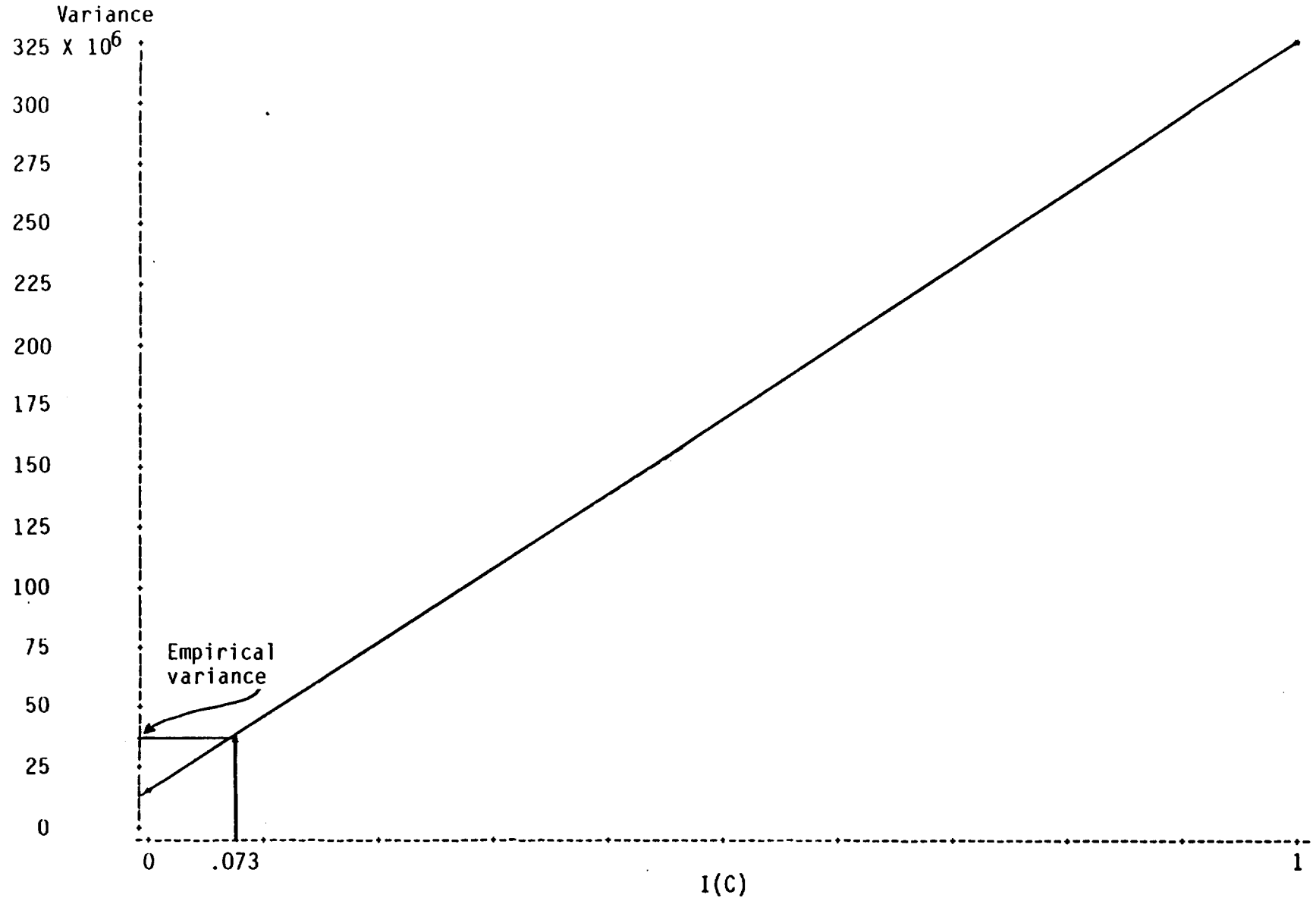

Figure 5-8.- Variance versus I(C) for hay cut in Beadle County.

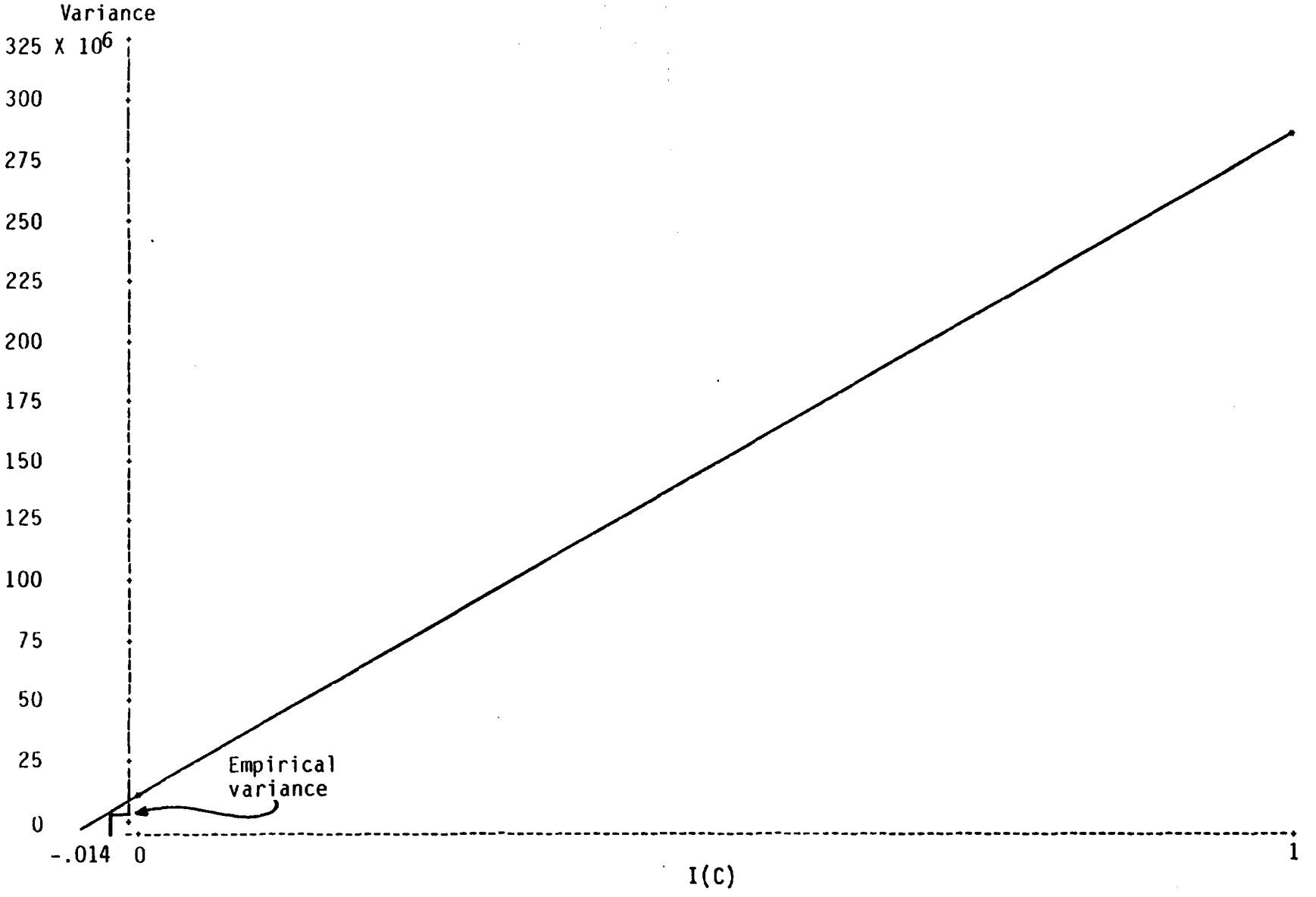

Figure 5-9.- Variance versus I(C) for flax in Beadle County.

5.1.3 BEHRENS-FISHER TEST

Table 5-7 contains the results of the Behrens-Fisher test described in section 4.5. This test is used as a guide in assessing the bias of the current county-level estimator. The corresponding confidence intervals for the estimated biases are in table 5-8.

This test, a significance test for the difference between the means of two normal populations, assumes that the two population variances are not the same. For a fixed crop and county, the eight estimates of hectarage associated with the eight training groups are considered as eight observations of a random variable Y_1.

For the same crop and county, the n sample segments of the 200 that fall in that county can each be thought of as providing an estimate of the mean hectarage per segment of that crop. By multiplying each of these n numbers by the total possible segments in the county, n estimates of the hectarage of the crop are obtained. Treating these as n observations of a random variable Y_2, this two-sample test can then be applied to test for the equality of means associated with the random variables Y_1 and Y_2. The following should be kept in mind in interpreting the test results: First, the eight observations of Y_1 are regression estimates based on means from training groups, some of which overlap; and second, the n observations of Y_2 arise from individual segments and thus produce a large sample variance. This variance will occur as part of the denominator in the test statistic, and it will likely produce a number which will fall within the interval determined by the critical values. This would imply that the hypothesis of equal means would not be rejected as often as might be expected, given that the efficacy of Y_2 as an estimator of the true population mean is suspect.

The other possibility for estimating the true mean was to use the segments falling within a county to obtain the direct expansion estimate of the true mean. This number would then be a constant C against which the mean from the distribution of Y_1 could be tested for equality. The difficulty with this possibility is that C is treated as the true mean, when in reality, it is only

TABLE 5-7.- BEHRENS-FISHER T-TEST OF MEAN ESTIMATES*

[α = .05]

County	1: Behrens-Fisher statistic 2: Critical values	Rangeland	Sunflowers	Corn	Wheat	Oats	Grass	Alfalfa	Hay cut	Flax
Beadle	1	-1.50	**	-1.60	†5.52	-1.47	†2.77	-1.12	-1.24	**
	2	±2.05	**	±2.02	±2.14	±2.03	±2.14	±2.06	±2.03	**
Clark	1	-1.20	-0.81	1.07	-1.26	†3.58	-.34	.72	-.55	1.87
	2	±2.04	±2.03	±2.03	±2.03	±2.06	±2.06	±2.06	±2.05	±2.11
Codington	1	.63	-.08	-.19	1.08	.57	-1.49	1.45	**	.10
	2	±2.07	±2.05	±2.06	±2.06	±2.09	±2.07	±2.07	**	±2.07
Hamlin	1	-1.80	**	†2.31	-.27	-1.52	†2.95	1.83	-.43	.59
	2	2.11	**	±2.11	±2.10	±2.12	±2.16	±2.12	±2.12	±2.12
Kingsbury	1	.20	-1.19	.08	-.87	†2.19	-1.46	.45	†3.60	.37
	2	±2.05	±2.04	±2.05	±2.05	±2.07	±2.05	±2.07	±2.19	±2.14
Spink	1	.73	-.55	-1.21	-.83	.52	.57	.24	.62	.29
	2	±2.04	±2.03	±2.01	±2.03	±2.10	±2.04	±2.13	±2.08	±2.08

*The hypothesis is that the population mean of the current county-level estimator equals the population mean of the direct expansion estimator.

†Hypothesis rejected.

**No crop present.

TABLE 5-8.- CONFIDENCE INTERVAL FOR ESTIMATED BIAS: CURRENT REGRESSION ESTIMATOR

[95% confidence]

County	Rangeland	Sunflowers	Corn	Wheat	Oats	Grass	Alfalfa	Hay cut	Flax
Beadle	-6.275		-3.015	7.368	-2.625	2.574	-2.065	-3.075	
	±8.573		±3.815	±2.861	±3.627	±1.987	±3.788	±5.011	
Clark	-4.867	-1.770	2.369	-2.879	4.067	-.739	1.024	-.729	1.808
	±8.233	±4.432	±4.487	±4.642	±2.346	±4.423	±2.922	±2.723	±2.034
Codington	2.641	-.090	-.466	1.777	1.255	-4.480	2.218		.294
	±8.607	±2.420	±4.968	±3.407	±4.568	±6.225	±3.161		±6.264
Hamlin	-9.222		8.533	-.682	-5.621	2.887	2.753	-.404	1.277
	±10.772		±7.792	±5.403	±7.817	±2.119	±3.196	±2.002	±4.568
Kingsbury	.809	-1.649	.267	-2.671	2.619	-4.376	.701	.928	.632
	±8.246	±2.825	±6.915	±6.300	±2.474	±6.137	±3.234	±.564	±3.683
Spink	2.311	-1.280	-2.344	-2.288	.531	.820	.431	.718	.195
	±6.474	±4.714	±3.907	±5.594	±2.147	±2.918	±3.891	±2.419	±1.427

an unbiased estimate of that mean, and it has a considerable amount of variance associated with it.

A decision to use the two-sample test was made, and, insofar as the mean of Y_2 can be considered the true population mean, the test results indicate that there is not enough statistical evidence to show that Y_1, the current county-level estimator, is biased.

5.1.4 ESTIMATION RESULTS FOR SOIL STRATUM 4

The current subanalysis district estimator was used to obtain crop hectarage estimates for soil stratum 4 for each of the eight training groups. (The Cardenas estimators were not evaluated on soil stratum 4 because their use requires knowing all of the land use stratum and soil stratum intersection means, which were not available.) In an analysis similar to that which was conducted for the six counties, an empirically derived value for I(C) was calculated. These results are shown in table 5-9. Again, the empirically observed values of I(C) cluster close to 0, with hay cut being the only exception. This gives additional credence to the conjecture that the variance formula with I(C) = 1 produces overestimates.

The Behrens-Fisher test described in section 5.1.3 was used to ascertain if the current estimator produced biased crop hectarage estimates on soil stratum 4. Table 5-10 gives the results of the Behrens-Fisher tests. No non-zero ground truth was present for flax or grass in the sample of 20 segments from soil stratum 4. Of the remaining seven crops, there was not enough statistical evidence to reject the null hypothesis of equal means. (A significant outcome for a crop would imply that the estimate for that crop is biased.)

5.2 RESULTS OF THE CARDENAS REGRESSION AND CARDENAS RATIO ESTIMATION PROCEDURES

5.2.1 COUNTY CROP PROPORTION AND COEFFICIENTS OF VARIATION

Tables 5-11 through 5-19 give, by crop, the proportion estimates and the "coefficients of variation" that were obtained for each training group for

TABLE 5-9.- THEORETICAL AND EMPIRICAL VARIANCE ESTIMATES
USING CURRENT USDA PROCEDURE FOR SOIL STRATUM 4

[Hectares x 10^4]

Item	Rangeland	Sunflowers	Corn	Wheat	Oats	Grass	Alfalfa	Hay cut	Flax
Average of variance estimates with I(C)=1	329.8	23.0	90.0	159.0	79.9	79.8	112.3	54.5	27.7
Empirical estimate of variance	59.7	.3	4.8	16.1	2.7	2.0	1.1	36.1	.8
Average of variance estimates with I(C)=0	14.5	.9	3.2	5.8	3.1	2.8	3.9	2.4	1.0
Empirically observed value of I(C)	.143	-.029	.018	.067	-.006	-.010	-.027	.646	-.010

TABLE 5-10.- BEHRENS-FISHER TEST OF MEAN ESTIMATES FOR SOIL STRATUM 4*

[α = 0.5]

Crop	Statistic	Critical value	Confidence interval for estimated bias	Relative bias
Rangeland	-0.408	±2.116	-2.134 ± 11.060	-0.088
Sunflowers	-.947	±2.094	-1.600 ± 3.539	-.948
Corn	-1.434	±2.096	-6.019 ± 8.800	-.456
Wheat	1.749	±2.113	5.096 ± 6.157	1.007
Oats	-.618	±2.104	-1.009 ± 3.433	-.350
Grass	†			
Alfalfa	-.420	±2.099	-.580 ± 2.902	-.181
Hay cut	.309	±2.159	.740 ± 5.161	-.202
Flax	†			

*The hypothesis is that the population mean of the Huddleston-Ray subanalysis district estimator equals the population mean of the direct expansion estimator.

†No crop present in sample from soil stratum 4.

TABLE 5-11.- COUNTY CROP PROPORTION AND COEFFICIENTS OF VARIATIONS FOR CARDENAS REGRESSION ESTIMATOR AND RATIO ESTIMATOR FOR RANGELAND

(a) Cardenas regression estimator

COUNTY HECTARES	6123	6146	6178	6247	6258	6345	6368	6567	MEAN PROPORTION C. V.	CROP HECTARES
BEADLE 322778	0.366 0.356	0.492 0.579	0.336 0.313	0.205 0.293	0.171 0.250	0.334 0.326	0.069 0.490	0.006 0.265	0.247 0.661	79821 2785651999
CLARK 248704	0.224 0.122	0.326 0.233	0.248 0.123	0.241 0.122	0.259 0.146	0.222 0.145	0.324 0.174	0.301 0.186	0.268 0.160	66678 113309463.36
CODINGTON 169126	0.160 0.157	0.323 0.537	0.239 0.171	0.259 0.182	0.319 0.198	0.119 0.087	0.203 0.113	0.573 0.495	0.274 0.511	46371 560693946.84
HAMLIN 130730	0.123 0.392	0.033 0.349	0.153 0.369	0.251 0.411	0.240 0.500	0.062 0.282	0.193 0.382	0.308 0.407	0.177 0.579	23098 178974077.64
KINGSBURY 213480	0.204 0.181	0.237 0.222	0.192 0.160	0.321 0.202	0.306 0.201	0.138 0.125	0.623 0.521	0.394 0.231	0.303 0.506	64593 1067119940
SPINK 359815	0.315 0.263	0.302 0.243	0.248 0.155	0.228 0.162	0.206 0.130	0.236 0.172	0.280 0.161	0.114 0.164	0.241 0.264	86769 524623698.98

(b) Cardenas ratio estimator

COUNTY HECTARES	6123	6146	6178	6247	6258	6345	6368	6567	MEAN PROPORTION C. V.	CROP HECTARES
BEADLE 322778	0.289 0.177	0.357 0.203	0.273 0.131	0.237 0.129	0.243 0.120	0.246 0.144	0.263 0.151	0.189 0.101	0.261 0.188	84255 250503014.86
CLARK 248704	0.289 0.185	0.359 0.205	0.293 0.147	0.219 0.120	0.254 0.135	0.227 0.154	0.251 0.142	0.185 0.108	0.260 0.208	64591 17969215.43
CODINGTON 169126	0.288 0.192	0.358 0.208	0.296 0.156	0.213 0.120	0.256 0.141	0.222 0.160	0.244 0.139	0.183 0.112	0.258 0.216	43591 88807531.98
HAMLIN 130730	0.269 0.222	0.323 0.227	0.232 0.182	0.250 0.166	0.212 0.128	0.198 0.153	0.211 0.125	0.184 0.128	0.235 0.192	30714 34797594.00
KINGSBURY 213480	0.291 0.184	0.358 0.223	0.248 0.124	0.262 0.146	0.232 0.121	0.253 0.151	0.285 0.178	0.196 0.103	0.265 0.180	56677 103881513.93
SPINK 359815	0.278 0.192	0.337 0.210	0.242 0.148	0.251 0.146	0.222 0.116	0.219 0.142	0.239 0.141	0.188 0.109	0.247 0.182	88882 262330548.79

TABLE 5-12.- COUNTY CROP PROPORTION AND COEFFICIENTS OF VARIATION FOR CARDENAS REGRESSION ESTIMATOR AND RATIO ESTIMATOR FOR FLAX

(a) Cardenas regression estimator

COUNTY HECTARES	6121	6146	6178	6247	6258	6345	6368	6567	MEAN PROPORTION C. V.	CROP HECTARES
BEAVER	0.004	0.004	0.011	0.023	0.014	0.004	0.014	0.023	0.012	3499
322778	0.321	0.360	0.423	0.825	0.349	0.449	0.434	0.973	0.678	6989782.84
CLARK	0.039	0.044	0.044	0.044	0.050	0.033	0.046	0.027	0.041	10169
248704	0.200	0.187	0.249	0.262	0.242	0.130	0.304	0.212	0.188	3646285.64
COMANCHE	0.068	0.069	0.068	0.064	0.073	0.058	0.056	0.050	0.063	10700
169126	0.227	0.205	0.427	0.260	0.293	0.181	0.226	0.220	0.121	1681540.27
HAMILTON	0.106	0.102	0.047	0.095	0.058	0.095	0.070	0.049	0.078	10159
130730	0.305	0.262	0.135	0.328	0.245	0.271	0.202	0.154	0.316	10333730.79
KINGSBURY	0.052	0.049	0.040	0.038	0.038	0.048	0.070	0.048	0.054	11454
213480	0.193	0.354	0.204	0.159	0.159	0.161	0.376	0.529	0.378	18766615.41
SPINK	0.008	0.014	0.008	0.015	0.016	0.015	0.014	0.030	0.015	5414
359415	0.349	0.266	0.206	0.391	0.404	0.341	0.320	0.535	0.461	6229748.79

(b) Cardenas ratio estimator

COUNTY HECTARES	6121	6146	6178	6247	6258	6345	6368	6567	MEAN PROPORTION C. V.	CROP HECTARES
BEAVER	0.044	0.060	0.045	0.045	0.046	0.042	0.051	0.036	0.047	15072
322778	0.208	0.243	0.269	0.221	0.227	0.154	0.274	0.242	0.152	5232111.64
CLARK	0.050	0.055	0.053	0.052	0.057	0.044	0.053	0.034	0.050	12377
248704	0.213	0.201	0.349	0.250	0.277	0.162	0.271	0.193	0.152	3523416.55
COMANCHE	0.052	0.054	0.055	0.050	0.061	0.046	0.053	0.033	0.051	8680
169126	0.221	0.187	0.364	0.265	0.293	0.170	0.266	0.174	0.166	2079426.95
HAMILTON	0.070	0.063	0.036	0.073	0.050	0.063	0.045	0.039	0.055	7165
130730	0.294	0.205	0.145	0.339	0.288	0.230	0.164	0.177	0.260	3478609.71
KINGSBURY	0.043	0.057	0.035	0.032	0.030	0.037	0.049	0.040	0.042	8883
213480	0.205	0.330	0.215	0.165	0.158	0.141	0.343	0.340	0.290	6639307.84
SPINK	0.069	0.059	0.037	0.058	0.044	0.053	0.047	0.039	0.050	17995
359415	0.254	0.229	0.158	0.282	0.242	0.195	0.212	0.217	0.199	12833948.84

TABLE 5-13.- COUNTY CROP PROPORTION AND COEFFICIENTS OF VARIATION FOR CARDENAS REGRESSION ESTIMATOR AND RATIO ESTIMATOR FOR HAY CUT

(a) Cardenas regression estimator

COUNTY HECTARES	G123	G146	G178	G247	G258	G345	G368	G567	MEAN PROPORTION C. V.	CROP HECTARES
HEADLE 322778	0.036 0.104	0.051 0.296	0.096 0.369	0.105 0.667	0.085 0.421	0.070 0.727	0.025 0.293	0.048 0.335	0.064 0.449	20795 87191599.84
CLARK 248704	0.020 0.153	0.027 0.253	0.036 0.240	0.036 0.466	0.007 0.119	0.039 0.579	0.034 0.211	0.060 0.429	0.032 0.476	8080 14809041.14
CODINGTON 169126	0.020 0.347	0.014 0.227	0.007 0.460	0.002 0.871	0.003 0.203	0.036 0.536	0.036 0.422	0.071 0.878	0.024 1.001	3981 15894196.41
HAMLIN 130730	0.033 0.299	0.023 0.253	0.022 0.439	0.025 0.225	-.003 0.269	0.046 0.417	0.047 0.525	0.057 0.509	0.031 0.605	4086 6116403.84
KINGSBURY 213480	0.020 0.214	0.013 0.169	0.018 0.269	0.021 0.236	0.011 0.216	0.027 0.275	0.023 0.337	0.045 0.350	0.022 0.464	4767 4895935.84
SPINK 359415	0.039 0.196	0.038 0.231	0.078 0.442	0.041 0.421	0.037 0.219	0.037 0.461	0.036 0.323	0.024 0.288	0.041 0.377	14859 31346629.41

(b) Cardenas ratio estimator

COUNTY HECTARES	G123	G146	G178	G247	G258	G345	G368	G567	MEAN PROPORTION C. V.	CROP HECTARES
HEADLE 322778	0.041 0.103	0.044 0.277	0.075 0.262	0.088 0.669	0.054 0.280	0.068 0.811	0.027 0.201	0.045 0.314	0.054 0.398	17443 48223096.21
CLARK 248704	0.029 0.077	0.054 0.340	0.090 0.246	0.129 0.847	0.054 0.254	0.092 1.029	0.030 0.193	0.050 0.377	0.066 0.523	16383 73322499.12
CODINGTON 169126	0.028 0.072	0.050 0.352	0.095 0.295	0.138 0.881	0.055 0.261	0.097 1.069	0.033 0.203	0.051 0.390	0.069 0.544	11671 40302708.12
HAMLIN 130730	0.030 0.179	0.024 0.131	0.056 0.464	0.009 0.071	0.072 0.735	0.017 0.085	0.047 0.584	0.026 0.140	0.035 0.603	4603 7697204.12
KINGSBURY 213480	0.035 0.165	0.033 0.189	0.056 0.247	0.041 0.327	0.051 0.321	0.041 0.393	0.019 0.183	0.041 0.248	0.039 0.284	8430 5721303.71
SPINK 359415	0.032 0.148	0.029 0.147	0.058 0.339	0.024 0.225	0.064 0.519	0.040 0.272	0.037 0.398	0.033 0.179	0.039 0.357	14006 24973061.07

TABLE 5-14.- COUNTY CROP PROPORTION AND COEFFICIENTS OF VARIATION FOR CARDENAS REGRESSION ESTIMATOR AND RATIO ESTIMATOR FOR ALFALFA

(a) Cardenas regression estimator

COUNTY HECTARES	G123	G146	G178	G247	G258	G345	G368	G567	MEAN PROPORTION C. V.	CROP HECTARES
BEADLE 322778	0.037 0.174	0.052 0.246	0.061 0.218	0.044 0.149	0.074 0.149	0.066 0.251	0.066 0.168	0.074 0.296	0.059 0.231	19107 19517815.36
CLARK 248704	0.060 0.169	0.070 0.199	0.049 0.116	0.057 0.131	0.074 0.172	0.062 0.175	0.061 0.159	0.034 0.113	0.058 0.214	14522 9655060.79
CODINGTON 169126	0.066 0.212	0.075 0.251	0.049 0.120	0.063 0.179	0.073 0.169	0.063 0.242	0.055 0.184	0.040 0.166	0.061 0.200	10235 4202062.55
HAMLIN 130730	0.079 0.260	0.040 0.219	0.058 0.128	0.056 0.136	0.081 0.176	0.070 0.184	0.092 0.214	0.051 0.134	0.071 0.205	9278 3608441.12
KINGSBURY 213480	0.076 0.229	0.084 0.162	0.069 0.141	0.066 0.154	0.110 0.232	0.062 0.149	0.067 0.150	0.049 0.100	0.076 0.235	16134 14422407.07
SPINK 359415	0.044 0.195	0.053 0.285	0.056 0.341	0.044 0.285	0.023 0.219	0.062 0.294	0.010 0.133	0.060 0.318	0.044 0.421	15795 44224477.41

(b) Cardenas ratio estimator

COUNTY HECTARES	G123	G146	G178	G247	G258	G345	G368	G567	MEAN PROPORTION C. V.	CROP HECTARES
BEADLE 322778	0.057 0.140	0.073 0.179	0.053 0.130	0.056 0.130	0.061 0.161	0.070 0.158	0.038 0.100	0.047 0.106	0.057 0.202	18307 13681456.70
CLARK 248704	0.055 0.143	0.069 0.190	0.053 0.137	0.054 0.131	0.056 0.145	0.062 0.145	0.039 0.110	0.047 0.111	0.054 0.163	13513 4444344.21
CODINGTON 169126	0.054 0.140	0.067 0.195	0.054 0.144	0.054 0.132	0.054 0.138	0.059 0.141	0.041 0.115	0.047 0.116	0.054 0.145	9075 1727134.21
HAMLIN 130730	0.060 0.169	0.071 0.161	0.066 0.173	0.054 0.135	0.059 0.137	0.066 0.149	0.046 0.110	0.057 0.133	0.060 0.131	7809 1041932.00
KINGSBURY 213480	0.061 0.144	0.078 0.188	0.052 0.129	0.059 0.141	0.067 0.190	0.081 0.181	0.034 0.098	0.046 0.106	0.060 0.267	12749 11582427.70
SPINK 359415	0.060 0.150	0.073 0.161	0.060 0.147	0.055 0.129	0.062 0.149	0.070 0.152	0.042 0.099	0.052 0.115	0.059 0.169	21348 12961666.86

TABLE 5-15.- COUNTY CROP PROPORTION AND COEFFICIENTS OF VARIATION FOR CARDENAS REGRESSION ESTIMATOR AND RATIO ESTIMATOR FOR GRASS

(a) Cardenas regression estimator

COUNTY HECTARES	G123	G146	G178	G247	G258	G345	G368	G567	MEAN PROPORTION C. V.	CROP HECTARES
BEADLE 322778	0.056 0.348	0.041 0.178	0.044 0.167	0.059 0.232	0.029 0.174	0.086 0.355	0.042 0.221	0.064 0.296	0.053 0.339	17164 33910986.21
CLARK 248704	0.124 0.542	0.097 0.436	0.054 0.224	0.048 0.138	0.026 0.114	0.079 0.219	0.072 0.202	0.036 0.112	0.067 0.486	16689 65725087.64
CODINGTON 169126	0.142 0.709	0.094 0.371	0.052 0.227	0.042 0.118	0.024 0.116	0.058 0.235	0.063 0.183	0.036 0.124	0.064 0.599	10886 42469631.98
HAMLIN 130730	0.044 0.263	0.044 0.313	0.052 0.207	0.055 0.186	0.045 0.207	0.073 0.298	0.075 0.181	0.053 0.162	0.068 0.286	8834 6379492.12
KINGSBURY 213480	0.101 0.279	0.101 0.275	0.064 0.224	0.127 0.413	0.138 0.536	0.137 0.514	0.041 0.184	-.013 0.241	0.086 0.636	18277 135258158.29
SPINK 359815	0.045 0.347	0.034 0.181	0.026 0.138	0.017 0.221	0.017 0.241	0.091 0.326	0.132 0.419	0.108 0.516	0.059 0.767	21094 261781136.12

(b) Cardenas ratio estimator

COUNTY HECTARES	G123	G146	G178	G247	G258	G345	G368	G567	MEAN PROPORTION C. V.	CROP HECTARES
BEADLE 322778	0.092 0.273	0.083 0.264	0.049 0.159	0.061 0.184	0.052 0.200	0.090 0.219	0.070 0.207	0.045 0.138	0.068 0.278	21877 37095179.70
CLARK 248704	0.086 0.280	0.075 0.260	0.045 0.168	0.052 0.169	0.044 0.183	0.081 0.206	0.063 0.193	0.039 0.130	0.061 0.301	15072 20515003.71
CODINGTON 169126	0.084 0.289	0.072 0.261	0.044 0.176	0.048 0.161	0.041 0.173	0.077 0.203	0.060 0.185	0.037 0.127	0.058 0.316	9792 9547763.84
HAMLIN 130730	0.101 0.320	0.083 0.272	0.051 0.202	0.048 0.142	0.041 0.150	0.088 0.196	0.069 0.171	0.041 0.121	0.065 0.356	8517 9196584.50
KINGSBURY 213480	0.099 0.277	0.095 0.278	0.055 0.155	0.076 0.205	0.064 0.223	0.103 0.239	0.080 0.228	0.053 0.151	0.078 0.254	16654 17844464.55
SPINK 359815	0.099 0.244	0.086 0.261	0.052 0.172	0.058 0.166	0.049 0.179	0.092 0.206	0.072 0.190	0.045 0.129	0.069 0.308	24871 58676291.84

TABLE 5-16.- COUNTY CROP PROPORTION AND COEFFICIENTS OF VARIATION FOR CARDENAS REGRESSION ESTIMATOR AND RATIO ESTIMATOR FOR OATS

(a) Cardenas regression estimator

COUNTY HECTARES	G123	G146	G178	G247	G258	G345	G368	G567	MEAN PROPORTION C. V.	CROP HECTARES
BEADLE 322778	0.064 0.244	0.003 0.146	0.059 0.526	0.045 0.258	0.064 0.267	0.033 0.298	0.053 0.324	0.059 0.354	0.047 0.442	15312 45807002.57
CLARK 248704	0.078 0.214	0.079 0.226	0.062 0.307	0.091 0.236	0.093 0.159	0.096 0.262	-.011 0.314	0.061 0.194	0.069 0.489	17237 71075836.79
CODINGTON 169126	0.083 0.291	0.117 0.333	0.102 0.302	0.118 0.252	0.106 0.186	0.097 0.347	0.082 0.188	0.043 0.187	0.094 0.260	15921 16865581.07
HAMLIN 130730	0.084 0.305	0.093 0.198	0.067 0.163	0.129 0.319	0.116 0.256	0.117 0.313	0.057 0.127	0.084 0.223	0.094 0.272	12235 11058740.57
KINGSBURY 213480	0.088 0.249	0.077 0.210	0.080 0.177	0.090 0.240	0.080 0.152	0.061 0.143	0.048 0.137	0.046 0.116	0.071 0.241	15228 13453116.84
SPINK 359815	0.052 0.315	0.020 0.092	0.064 0.170	0.055 0.241	0.061 0.276	0.039 0.164	0.110 0.442	0.046 0.192	0.056 0.466	20076 87577411.55

(b) Cardenas ratio estimator

COUNTY HECTARES	G123	G146	G178	G247	G258	G345	G368	G567	MEAN PROPORTION C. V.	CROP HECTARES
BEADLE 322778	0.075 0.178	0.071 0.205	0.073 0.155	0.090 0.228	0.081 0.143	0.068 0.170	0.055 0.150	0.050 0.126	0.070 0.167	22718 18117076.98
CLARK 248704	0.073 0.175	0.073 0.250	0.070 0.161	0.092 0.234	0.082 0.142	0.068 0.176	0.051 0.135	0.053 0.143	0.070 0.194	17419 11377703.71
CODINGTON 169126	0.072 0.177	0.073 0.263	0.070 0.168	0.093 0.239	0.082 0.144	0.068 0.184	0.050 0.129	0.055 0.153	0.070 0.197	11893 5492469.71
HAMLIN 130730	0.071 0.196	0.070 0.173	0.087 0.211	0.101 0.264	0.092 0.157	0.083 0.227	0.050 0.117	0.070 0.191	0.078 0.206	10193 4422737.14
KINGSBURY 213480	0.078 0.196	0.069 0.189	0.075 0.153	0.087 0.241	0.080 0.156	0.068 0.169	0.061 0.184	0.043 0.105	0.070 0.194	14964 8393744.84
SPINK 359815	0.074 0.183	0.070 0.167	0.082 0.180	0.096 0.240	0.087 0.146	0.077 0.195	0.054 0.134	0.060 0.156	0.075 0.186	25894 24924206.57

TABLE 5-17.- COUNTY CROP PROPORTION AND COEFFICIENTS OF VARIATION FOR CARDENAS REGRESSION ESTIMATOR AND RATIO ESTIMATOR FOR WHEAT

(a) Cardenas regression estimator

COUNTY HECTARES	G123	G146	G178	G247	G258	G345	G368	G567	MEAN PROPORTION C. V.	CROP HECTARES
BEADLE 322778	0.056 0.316	0.098 0.250	0.040 0.377	0.047 0.324	0.041 0.319	0.052 0.363	0.030 0.223	0.075 0.507	0.055 0.398	17681 49492259.55
CLARK 248704	0.083 0.088	0.153 0.163	0.170 0.191	0.149 0.159	0.151 0.152	0.121 0.144	0.163 0.214	0.199 0.210	0.149 0.233	36457 73867950.84
CODINGTON 169126	0.116 0.133	0.165 0.193	0.170 0.206	0.162 0.198	0.160 0.181	0.131 0.166	0.027 0.128	0.187 0.196	0.140 0.363	23669 73721837.55
HAMLIN 130730	0.211 0.241	0.231 0.231	0.201 0.186	0.174 0.152	0.162 0.145	0.197 0.174	0.216 0.184	0.279 0.201	0.209 0.172	27302 22168643.64
KINGSBURY 213480	0.151 0.162	0.174 0.181	0.170 0.150	0.164 0.149	0.213 0.210	0.140 0.125	0.253 0.277	0.279 0.228	0.193 0.261	41210 115472485.43
SPINK 359415	0.058 0.182	0.090 0.306	0.104 0.211	0.067 0.228	0.057 0.301	0.072 0.213	0.058 0.171	0.091 0.290	0.075 0.242	26912 42552967.93

(b) Cardenas ratio estimator

COUNTY HECTARES	G123	G146	G178	G247	G258	G345	G368	G567	MEAN PROPORTION C. V.	CROP HECTARES
BEADLE 322778	0.085 0.099	0.116 0.127	0.111 0.132	0.103 0.127	0.099 0.138	0.095 0.153	0.092 0.137	0.146 0.208	0.106 0.180	34175 37995587.55
CLARK 248704	0.089 0.114	0.114 0.136	0.106 0.135	0.096 0.119	0.094 0.136	0.103 0.211	0.090 0.133	0.128 0.177	0.103 0.135	25593 12019924.27
CODINGTON 169126	0.091 0.121	0.121 0.143	0.105 0.139	0.093 0.116	0.092 0.138	0.106 0.228	0.091 0.134	0.121 0.163	0.102 0.125	17305 4692260.29
HAMLIN 130730	0.094 0.124	0.151 0.175	0.122 0.167	0.103 0.123	0.106 0.160	0.091 0.114	0.115 0.154	0.132 0.150	0.114 0.177	14933 6947933.43
KINGSBURY 213480	0.078 0.090	0.110 0.120	0.117 0.135	0.112 0.147	0.106 0.148	0.084 0.112	0.093 0.150	0.173 0.256	0.109 0.268	23305 38933524.86
SPINK 359415	0.088 0.106	0.135 0.147	0.119 0.145	0.105 0.123	0.105 0.146	0.090 0.106	0.106 0.143	0.145 0.181	0.112 0.162	40188 53360705.14

TABLE 5-18.- COUNTY CROP PROPORTION AND COEFFICIENTS OF VARIATION FOR CARDENAS REGRESSION ESTIMATOR AND RATIO ESTIMATOR FOR CORN

(a) Cardenas regression estimator

COUNTY HECTARES	G123	G146	G178	G247	G258	G345	G368	G567	MEAN PROPORTION C. V.	CROP HECTARES
BEADLE 322778	0.079 0.228	0.090 0.255	0.119 0.228	0.043 0.183	0.052 0.181	0.071 0.192	0.121 0.359	0.104 0.268	0.085 0.342	27470 88256701.55
CLARK 248704	0.137 0.127	0.127 0.137	0.146 0.144	0.139 0.149	0.157 0.135	0.123 0.113	0.143 0.131	0.110 0.091	0.135 0.109	33609 13444108.50
CODINGTON 169126	0.157 0.147	0.149 0.152	0.160 0.163	0.164 0.167	0.185 0.158	0.128 0.120	0.142 0.139	0.105 0.088	0.149 0.163	25147 16842437.36
HAMLIN 130730	0.292 0.226	0.208 0.180	0.290 0.224	0.230 0.196	0.335 0.236	0.312 0.286	0.279 0.218	0.223 0.165	0.271 0.168	35457 35625761.98
KINGSBURY 213480	0.226 0.190	0.158 0.161	0.227 0.216	0.233 0.188	0.297 0.217	0.247 0.229	0.204 0.176	0.161 0.125	0.219 0.208	46769 94412921.07
SPINK 359815	0.063 0.252	0.108 0.273	0.106 0.246	0.052 0.258	0.043 0.215	0.075 0.287	0.119 0.379	0.094 0.252	0.082 0.343	29678 10397962.84

(b) Cardenas ratio estimator

COUNTY HECTARES	G123	G146	G178	G247	G258	G345	G368	G567	MEAN PROPORTION C. V.	CROP HECTARES
BEADLE 322778	0.165 0.141	0.137 0.127	0.180 0.173	0.156 0.139	0.195 0.158	0.164 0.154	0.167 0.139	0.133 0.100	0.162 0.126	52380 43426553.27
CLARK 248704	0.167 0.149	0.140 0.139	0.172 0.165	0.151 0.148	0.189 0.159	0.161 0.168	0.165 0.139	0.132 0.105	0.160 0.115	39686 20843769.27
CODINGTON 169126	0.168 0.155	0.143 0.145	0.171 0.164	0.148 0.152	0.187 0.162	0.162 0.178	0.167 0.142	0.133 0.109	0.160 0.110	27040 8888850.12
HAMLIN 130730	0.182 0.169	0.160 0.146	0.207 0.180	0.143 0.147	0.200 0.174	0.205 0.218	0.212 0.164	0.167 0.130	0.185 0.138	24128 11030848.27
KINGSBURY 213480	0.162 0.141	0.131 0.126	0.188 0.195	0.164 0.143	0.203 0.170	0.165 0.139	0.165 0.147	0.132 0.099	0.164 0.149	34968 27065905.27
SPINK 359815	0.174 0.150	0.149 0.130	0.198 0.174	0.151 0.136	0.200 0.162	0.188 0.182	0.193 0.148	0.153 0.113	0.176 0.125	63238 62842461.27

TABLE 5-19.- COUNTY CROP PROPORTION AND COEFFICIENTS OF VARIATION FOR CARDENAS REGRESSION ESTIMATOR AND RATIO ESTIMATOR FOR SUNFLOWERS

(a) Cardenas regression estimator

COUNTY HECTARES	G123	G146	G178	G247	G258	G345	G368	G567	MEAN PROPORTION C. V.	CROP HECTARES
BEADLE 322778	0.048 0.207	0.050 0.275	0.060 0.242	0.085 0.612	0.036 0.207	0.042 0.209	0.027 0.234	0.058 0.300	0.051 0.349	16416 32742077.07
CLARK 248704	0.034 0.154	0.044 0.210	0.036 0.163	0.088 0.554	0.048 0.334	0.056 0.251	0.023 0.136	0.036 0.167	0.046 0.434	11395 24442359.27
CODINGTON 169126	0.040 0.169	0.047 0.233	0.046 0.192	0.095 0.614	0.043 0.214	0.050 0.215	0.021 0.109	0.044 0.195	0.048 0.430	8165 12315819.70
HAMLIN 130730	0.057 0.203	0.067 0.241	0.060 0.231	0.159 0.819	0.050 0.252	0.066 0.248	0.023 0.099	0.074 0.251	0.072 0.545	9403 26276110.55
KINGSBURY 213480	0.055 0.243	0.049 0.421	0.071 0.346	0.054 0.444	0.028 0.237	0.027 0.195	0.033 0.270	0.064 0.419	0.048 0.351	10144 12673518.86
SPINK 359415	0.030 0.671	0.021 0.408	-.011 0.121	0.050 0.765	0.065 1.020	0.049 0.639	0.025 0.289	0.013 0.419	0.030 0.804	10825 75757934.70

(b) Cardenas ratio estimator

COUNTY HECTARES	G123	G146	G178	G247	G258	G345	G368	G567	MEAN PROPORTION C. V.	CROP HECTARES
BEADLE 322778	0.031 0.299	0.029 0.254	0.012 0.077	0.051 0.344	0.050 0.554	0.046 0.357	0.027 0.200	0.026 0.209	0.034 0.398	10970 19099918.41
CLARK 248704	0.034 0.362	0.025 0.216	0.012 0.078	0.055 0.394	0.048 0.491	0.039 0.304	0.022 0.172	0.028 0.235	0.033 0.426	8168 12105383.84
CODINGTON 169126	0.036 0.396	0.023 0.195	0.012 0.080	0.057 0.424	0.048 0.460	0.037 0.276	0.021 0.157	0.029 0.250	0.033 0.452	5553 6287232.12
HAMLIN 130730	0.050 0.466	0.025 0.154	0.015 0.085	0.079 0.488	0.060 0.409	0.039 0.222	0.021 0.128	0.038 0.285	0.041 0.520	5356 7761323.98
KINGSBURY 213480	0.025 0.201	0.036 0.314	0.013 0.082	0.043 0.282	0.051 0.671	0.055 0.444	0.033 0.249	0.024 0.182	0.035 0.416	7444 9571024.70
SPINK 359415	0.040 0.377	0.028 0.207	0.014 0.079	0.065 0.407	0.056 0.477	0.045 0.292	0.025 0.165	0.033 0.241	0.038 0.436	13790 36189867.84

each county. For example, in Beadle County the proportion estimate of rangeland using the training group G123 is .366 of the total hectares in the three strata over which the estimate was obtained. The corresponding "coefficient of variation" is .356 and is defined as the ratio of the square root of the variance, calculated by the formula, and the average of the hectarage estimates across the eight training groups. The next to the last column contains, for each county, the mean proportion estimate of rangeland and the sample coefficient of variation. In the last column are the estimates represented as hectares.

By comparing the "coefficients of variation" that were computed using the formula variance for a training group to the sample coefficient of variation, one can see that the variance formula of $\hat{Y}$, which was derived under the assumption that the within-county variance is equal for all counties, seems to underestimate the true variance.

5.2.2 BEHRENS-FISHER TEST

In tables 5-20 and 5-21, the same two sample tests used to evaluate the bias of the current county regression estimator was used to test for bias in the two Cardenas estimators. The corresponding confidence intervals for the estimated biases are in tables 5-22 and 5-23. The same caution encouraged in examining the results in the first application of the test is advised here also. Because the hypothesis for the Cardenas ratio estimator is rejected for 10 crop-county combinations and the hypothesis for the Cardenas regression estimator is rejected for 8 combinations, both estimators can probably be considered biased by this test.

5.2.3 F-TESTS OF VARIANCE

The two-sided F-test was used to provide some idea of how the variance of the current county-level estimator compared to the variances of the two Cardenas estimators and how the variances of the Cardenas estimators compared to each other. These tests cannot be appealed to unequivocally because one of the assumptions for performing the test is that the samples are from independent

TABLE 5.20.- BEHRENS-FISHER T-TEST OF MEAN ESTIMATES: CARDENAS REGRESSION ESTIMATOR*

[α = .05]

County	1: Behrens-Fisher statistic 2: Critical values	Rangeland	Sunflowers	Corn	Wheat	Oats	Grass	Alfalfa	Hay cut	Flax
Beadle	1	†-2.40	**	-1.02	0.68	-1.21	†3.07	-0.77	-0.99	**
	2	±2.18	**	±2.06	±2.08	±2.04	±2.10	±2.03	±2.04	**
Clark	1	-.46	-.38	.92	.08	2.12	-.26	.07	-.31	1.08
	2	±2.05	±2.04	±2.04	±2.06	±2.14	±2.07	±2.04	±2.05	±2.04
Codington	1	1.86	1.17	-.38	†2.25	-.07	-1.41	.82	**	-.25
	2	±2.17	±2.09	±2.07	±2.16	±2.07	±2.08	±2.06	**	±2.05
Hamlin	1	-.46	**	1.94	†2.63	-1.57	†3.59	1.68	1.07	.09
	2	±2.14	**	±2.12	±2.12	±2.11	±2.15	±2.11	±2.15	±2.12
Kingsbury	1	1.94	.75	-.06	1.18	1.49	-1.36	.76	†4.13	-.63
	2	±2.18	±2.06	±2.07	±2.08	±2.08	±2.09	±2.06	±2.23	±2.07
Spink	1	.29	†-2.12	-.82	†-2.88	1.19	1.38	-.80	.07	-.07
	2	±2.08	±2.03	±2.05	±2.02	±2.12	±2.13	±2.04	±2.05	±2.03

*The hypothesis is that the population mean of the current county-level estimator equals population mean of direct expansion estimator.

†Hypothesis rejected.

**No crop present.

TABLE 5-21.- BEHRENS-FISHER T-TEST OF MEAN ESTIMATES: CARDENAS RATIO ESTIMATOR*

[α = .05]

County	1: Behrens-Fisher statistic 2: Critical values	Rangeland	Sunflowers	Corn	Wheat	Oats	Grass	Alfalfa	Hay cut	Flax
Beadle	1	†-2.96	**	1.53	†3.59	-0.40	†4.13	-0.86	-1.27	**
	2	±2.04	**	±2.04	±2.06	±2.03	±2.10	±2.03	±2.03	**
Clark	1	-.58	-0.76	1.62	-1.21	†2.55	-.47	-.12	1.17	1.76
	2	±2.06	±2.03	±2.04	±2.03	±2.05	±2.04	±2.04	±2.12	±2.04
Codington	1	2.02	.39	-.08	1.24	-.81	-1.60	.53	**	-.51
	2	±2.08	±2.08	±2.06	±2.06	±2.06	±2.06	±2.06	**	±2.05
Hamlin	1	.22	**	.52	.41	-1.86	†3.28	1.21	1.29	-.60
	2	±2.11	**	±2.11	±2.11	±2.10	±2.17	±2.10	±2.16	±2.11
Kingsbury	1	1.91	.18	-1.11	-.50	1.45	-1.62	.10	†7.4	-1.18
	2	±2.06	±2.06	±2.05	±2.06	±2.06	±2.05	±2.06	±2.24	±2.05
Spink	1	.42	-1.92	†2.21	-1.99	†2.65	2.08	-.16	-.07	†3.41
	2	±2.05	±2.02	±2.03	±2.02	±2.05	±2.05	±2.02	±2.04	±2.05

*The hypothesis is that the population mean of the current county-level estimator equals population mean of direct expansion estimator.

†Hypothesis rejected.

**No crop present.

TABLE 5-22.- CONFIDENCE INTERVAL FOR ESTIMATED BIAS: CARDENAS REGRESSION ESTIMATOR

[95% confidence]

County	Rangeland	Sunflowers	Corn	Wheat	Oats	Grass	Alfalfa	Hay cut	Flax
Beadle	-13.106 ±11.887		-2.034 ±4.106	0.792 ±2.430	-2.196 ±3.701	2.634 ±1.796	-1.335 ±3.576	-2.488 ±5.147	
Clark	-1.897 ±8.432	-0.842 ±4.541	2.037 ±4.530	.184 ±4.971	2.814 ±2.840	-.571 ±4.524	.098 ±2.834	-.412 ±2.710	0.913 ±1.732
Codington	9.638 ±11.212	1.476 ±2.651	-.927 ±5.085	4.485 ±4.300	-.185 ±4.401	-4.327 ±6.353	1.234 ±3.116		-.754 ±6.097
Hamlin	-2.552 ±11.911		7.410 ±8.078	7.045 ±5.680	-5.699 ±7.627	3.430 ±2.058	2.463 ±3.107	1.072 ±2.152	.191 ±4.541
Kingsbury	10.238 ±11.502	1.071 ±2.956	-.203 ±7.272	3.856 ±6.780	1.805 ±2.516	-4.367 ±6.716	1.174 ±3.183	1.256 ±6.793	-.957 ±3.144
Spink	.969 ±6.992	-4.939 ±4.739	-1.663 ±4.144	-7.761 ±5.448	1.269 ±2.265	2.373 ±3.679	-1.234 ±3.161	.076 ±2.277	-.045 ±1.281

TABLE 5-23.- CONFIDENCE INTERVAL FOR ESTIMATED BIAS: CARDENAS RATIO ESTIMATOR

[95% confidence]

County	Rangeland	Sunflowers	Corn	Wheat	Oats	Grass	Alfalfa	Hay cut	Flax
Beadle	-12.216 ±8.440		2.963 ±3.953	4.101 ±2.364	-0.710 ±3.597	3.579 ±1.820	-1.515 ±3.553	-3.160 ±5.043	
Clark	-2.440 ±8.613	-1.682 ±4.478	3.619 ±4.567	-2.774 ±4.674	2.861 ±2.300	-.992 ±4.287	-.165 ±2.794	1.750 ±3.172	1.488 ±1.730
Codington	8.574 ±8.841	.476 ±2.534	-.203 ±5.006	2.049 ±3.397	-1.686 ±4.270	-4.746 ±6.088	.790 ±3.076		-1.527 ±6.100
Hamlin	1.140 ±10.890	1.919 ±7.829	1.049 ±5.460	-6.689 ±7.557	3.276 ±2.166	1.751 ±3.040	1.323 ±2.210	-1.259 ±4.418	
Kingsbury	7.837 ±8.449	.252 ±2.923	-3.782 ±6.975	-1.574 ±6.415	1.725 ±2.452	-4.860 ±6.142	.148 ±3.154	2.367 ±.717	-1.737 ±3.021
Spink	1.350 ±6.564	-4.406 ±4.646	4.376 ±4.033	-5.372 ±5.470	2.496 ±1.934	3.053 ±3.005	-.234 ±3.050	-.077 ±2.246	2.218 ±1.336

populations. The overlapping training groups obviate satisfying the independence criterion; however, keeping in mind this limitation, the results in table 5-24 indicate that whenever there is a significant difference in variances, the Cardenas regression estimator and the Cardenas ratio estimator appear to have larger variances than the current regression estimator. Additionally, there seems to be no significant difference in variances of the Cardenas regression and the Cardenas ratio estimators.

5.2.4 RESULTS OF THE CLASSY-BASED DIRECT PROPORTION ESTIMATION PROCEDURE

The CLASSY-based direct proportion estimation procedure using the maximum likelihood approach or the least squares approach can be outlined as follows:

1. Apply the CLASSY clustering algorithm (noncrop specific to the county of interest, sampled 1/64) to estimate m, $p(x|i)$, and α_i.
2. Randomly choose 500 labeled pixels from the segments within the county of interest.
3. Use the maximum likelihood approach described in section 2.3.1 to estimate $\beta_{\ell i}$.
4. Use least squares approach described in section 2.3.2 to estimate $\beta_{\ell i}$.
5. Compute the proportion of class ℓ as $\hat{\pi}_\ell = \sum_{i=1}^{m} \alpha_i \hat{\beta}_{\ell i}$
6. Repeat steps 2, 3, 4, and 5, 50 times to estimate the bias and MSE of $\hat{\pi}_\ell$.

The procedure was also carried out without using the maximum likelihood approach or the least squares approach; the 500 labeled pixels were used to compute the proportion π_ℓ directly. This is referred to as the simple random sample approach. The number of labeled pixels in each county are as follows: Beadle, 8442; Clark, 7264; Codington, 5430; Hamlin, 3588; Kingsbury, 6086; and Spink, 9480.

TABLE 5-24.- F-TESTS OF VARIANCE

[Critical values: $F_{0.01;7,7} = 0.143$, $F_{0.99;7,7} = 6.99$]

County	Hypothesis*	Rangeland	Sunflowers	Corn	Wheat	Oats	Grass	Alfalfa	Hay cut	Flax
Beadle	H_1	†8.37	†31.16	†20.97	0.38	1.77	0.57	0.25	2.37	1.35
	H_2	.75	†18.18	†10.32	.29	.70	.62	.18	1.31	1.01
	H_3	†11.12	1.71	2.03	1.30	2.53	.91	1.43	1.81	1.34
Clark	H_1	2.75	†7.67	2.70	†13.41	4.45	1.42	.47	.91	†.13
	H_2	4.36	3.80	4.18	2.18	.71	.44	.24	4.48	†.12
	H_3	.63	2.02	.65	6.15	6.25	3.20	1.99	.20	1.03
Codington	H_1	†11.65	†18.83	3.31	†13.67	.53	1.61	.60	2.90	†.08
	H_2	1.85	†9.61	1.75	.87	.17	.36	.25	†7.36	†.09
	H_3	6.31	1.96	1.89	†15.71	3.07	4.45	2.43	.39	.81
Hamlin	H_1	†9.47	†179.90	4.75	6.98	.38	.80	.51	2.70	.87
	H_2	1.84	†53.14	1.47	2.20	.15	1.16	.15	3.39	.29
	H_3	5.14	3.39	3.23	3.17	2.50	.69	3.46	.79	2.97
Kingsbury	H_1	†21.38	†21.74	6.85	†7.38	1.33	†7.94	.73	1.85	.24
	H_2	2.08	†16.42	1.96	2.49	.83	1.05	.59	2.17	†.09
	H_3	†10.27	1.32	3.49	2.97	1.60	†7.58	1.25	.86	2.83
Spink	H_1	2.50	1.16	6.02	.37	1.37	†7.36	.16	.51	.26
	H_2	1.25	.56	3.64	.46	.39	1.65	†.05	.41	.53
	H_3	2.00	2.09	1.65	.80	3.51	4.46	3.41	1.26	.49

*H_1: Variance Cardenas regression estimator equals variance current regression estimator.

H_2: Variance Cardenas ratio estimator equals variance current regression estimator.

H_3: Variance Cardenas regression estimator equals variance Cardenas ratio estimator.

†Hypothesis rejected.

5.2.5 STATISTICS FOR DIRECT PROPORTION ESTIMATORS

In tables 5-25 and 5-26 are some statistics that were calculated for each of the direct proportion estimators. For each crop, the bias, mean squared error, and F-ratio are listed. The bias is the difference between the average of the 50 proportion estimates of a crop given by the particular estimator and the proportion of ground truth of that crop in the sample segments that are in the county.

The mean squared error (MSE) is the average of the squared errors over the 50 runs. The F-ratio is the ratio of the variance of the 50 estimates given by the direct proportion estimator to the variance of the 50 estimates obtained, each using a simple random sample of 500 labeled pixels to compute the proportion directly.

There seem to be no significant differences in variances of the CLASSY direct-proportion estimators using the simple random sample approach and either the maximum likelihood or the least squares approach.

5.2.6 RELATIVE BIASES OF ALTERNATIVE COUNTY ESTIMATORS

Table 5-27 lists the relative biases of each of the alternative county-level estimators considered. The relative bias is defined by the equation:

$$\text{Relative bias} = \frac{\text{estimator proportion of crop - true proportion of crop}}{\text{true proportion of crop}}$$

The true proportion of a crop in a county is declared to be the proportion of that crop determined by all of the ground-truth pixels in the sample of segments from that county. It can be seen that the Cardenas estimators produce relative biases consistently larger than those of the direct proportion estimators, which appear to have about a 10-percent relative bias.

5.3 STUDY RESULTS: PREPROCESSING

The preprocessing study was to determine if any of three candidate preprocessors might improve crop area estimation at a county level by correcting for

TABLE 5-25.- BIAS, MEAN SQUARED ERROR, AND F-RATIO USING THE MAXIMUM LIKELIHOOD APPROACH

[Critical values: $F_{0.01;49,49} = 0.511$, $F_{0.99;49,49} = 1.97$]

County	Grass			Alfalfa			Hay Cut			Flax			Other		
	Bias	Mean squared error	F-ratio	Bias	Mean squared error	F-ratio	Bias	Mean squared error	F-ratio	Bias	Mean squared error	F-ratio	Bias	Mean squared error	F-ratio
Beadle	-0.0003	0.0000	1.15	0.0119	0.0003	1.20	0.0010	0.0001	1.10	No crop present			0.0137	0.0042	1.29
Clark	-.0023	.0001	.96	-.0013	.0001	.95	-.0001	.0001	.82	-0.0016	0.0000	1.06	.0039	.0004	1.35
Codington	-.0036	.0002	.88	.0067	.0001	1.08	No crop present			.0008	.0001	1.09	.0125	.0006	1.12
Hamlin	-.0056	.0001	*2.43	.0480	.0027	*7.67	.0216	.0006	*8.12	-.0049	.0003	*3.91	-.0066	.0013	*3.75
Kingsbury	.0108	.0003	.89	-.0062	.0001	.70	-.0010	.0000	1.55	-.0031	.0001	.98	.0108	.0003	1.08
Spink	-.0024	.0000	.82	.0006	.0001	1.04	-.0013	.0001	.92	.0017	.0000	1.18	.0041	.0003	.60
	Rangeland			Corn			Wheat			Oats			Sunflower		
Beadle	-0.0400	0.0018	0.75	0.0123	0.0003	0.69	-0.0006	0.0001	1.06	0.0020	0.0001	0.97	No crop present		
Clark	-.0035	.0004	.95	.0139	.0003	.83	-.0130	.0003	.69	-.0017	.0000	.86	0.0058	0.0002	1.31
Codington	-.0079	.0002	.98	-.0256	.0008	.66	.0070	.0002	1.15	.0109	.0003	.96	-.0009	.0000	1.12
Hamlin	-.1439	.0207	*.31	.0701	.0061	*5.44	-.0328	.0013	1.93	.0542	.0040	*7.21	No crop present		
Kingsbury	.0074	.0003	1.03	-.0200	.0005	.54	.0052	.0002	.91	-.0003	.0001	1.09	-.0036	.0001	.99
Spink	-.0097	.0004	.90	.0073	.0002	.68	.0021	.0002	.83	.0004	.0001	1.05	-.0027	.0001	.55

*Hypothesis rejected.

TABLE 5-26.- BIAS, MEAN SQUARED ERROR, AND F-RATIO USING THE LEAST SQUARES APPROACH

[Critical values: $F_{0.01;49,49} = 0.511$, $F_{0.99;49,49} = 1.97$]

County	Grass			Alfalfa			Hay Cut			Flax			Other		
	Bias	Mean squared error	F-ratio	Bias	Mean squared error	F-ratio	Bias	Mean squared error	F-ratio	Bias	Mean squared error	F-ratio	Bias	Mean squared error	F-ratio
Beadle	-0.0007	0.0000	0.92	0.0084	0.0002	1.23	-0.0007	0.0001	1.07	No crop present			0.0150	0.0004	1.24
Clark	-.0044	.0001	.95	-.0015	.0001	1.05	-.0034	.0000	.76	-0.0031	0.0000	1.04	.0079	.0004	1.38
Codington	-.0021	.0001	1.01	.0060	.0000	1.09	No crop present			.0035	.0001	1.19	.0145	.0006	1.08
Hamlin	-.0161	.0003	*2.08	.0388	.0019	*7.11	.0051	.0002	*6.81	-.0223	.0011	*7.18	-.0722	.0071	*5.68
Kingsbury	.0018	.0001	.92	-.0084	.0001	.71	.0002	.0000	.99	-.0011	.0001	1.02	.0225	.0007	1.14
Spink	-.0037	.0000	.73	.0010	.0001	1.43	-.0029	.0000	.90	.0014	.0000	1.14	.0040	.0003	1.00
	Rangeland			Corn			Wheat			Oats			Sunflower		
Beadle	-0.0418	0.0019	0.66	0.0183	0.0004	0.68	0.0000	0.0000	1.03	0.0015	0.0001	0.94	No crop present		
Clark	-.0086	.0004	1.02	.0159	.0004	1.08	-.0101	.0002	.61	-.0015	.0000	.79	0.0092	0.0002	1.52
Codington	-.0074	.0002	1.03	-.0370	.0014	.56	.0102	.0002	1.28	.0142	.0004	1.16	-.0019	.0000	1.03
Hamlin	-.1729	.0306	*2.96	.1547	.0262	*10.48	-.0181	.0010	*5.09	.1030	.0122	*10.36	No crop present		
Kingsbury	-.0063	.0002	.85	-.0191	.0005	.65	.0152	.0003	.96	.0002	.0001	1.12	-.0052	.0000	.80
Spink	-.0115	.0004	1.02	.0206	.0006	1.08	.0032	.0001	.80	.0017	.0000	1.18	-.0139	.0002	.51

*Hypothesis rejected

TABLE 5-27.- RELATIVE BIAS OF ALTERNATIVE COUNTY ESTIMATORS

County	Estimator	Rangeland	Sunflowers	Corn	Wheat	Oats	Grass	Alfalfa	Hay cut	Flax
Beadle	*1	-0.450		-0.270	0.288	-0.417	3.250	-0.261	-0.374	
	†2	-.420		.393	1.489	-.135	4.422	-.292	-.475	
	**3	-.090		.106	-.014	.025	-.024	.149	.009	
	††4	-.094		.158	.000	.019	-.057	.106	-.007	
	§5	-.216		-.400	2.675	-.498	3.181	-.398	-.462	
Clark	1	-.099	-.221	.303	.020	1.681	-.116	0.027	-.164	0.526
	2	-.127	-.442	.539	-.294	1.709	-.202	-.045	.696	.858
	3	-.012	.099	.135	-.089	-.066	-.031	-.023	-.003	-.061
	4	-.029	.158	.154	-.069	-.058	-.059	-.026	-.008	-.117
	5	-.253	-.465	.353	-.305	2.430	-.150	.278	-.290	1.042
Codington	1	1.188	.894	-.088	.980	-.030	-.509	.460		-.155
	2	1.057	.289	-.019	.448	-.270	-.559	.294		-.315
	3	-.064	-.036	-.159	.101	.115	-.028	.164		.011
	4	-.060	-.076	-.230	.147	.109	-.016	.147		.047
	5	.325	-.054	-.044	.388	.201	-.527	.826		.061
Hamlin	1	-.186		.758	1.138	-.490	4.025	1.211	1.181	.040
	2	.083		.196	.170	-.575	3.844	.861	1.456	-.266
	3	-.670		.456	-.345	.301	-.403	1.500	1.612	-.066
	4	-.805		1.010	-.191	.573	-1.158	1.213	.381	-.299
	5	-.671		.873	-.110	-.483	3.388	1.353	-.445	.270
Kingsbury	1	1.095	.534	-.014	.446	.642	-.441	.316	6.597	-.216
	2	.838	.126	-.263	-.182	.613	-.490	.040	12.434	-.393
	3	.052	-.118	-.091	.039	-.007	.071	-.109	-.333	-.046
	4	-.044	-.171	-.087	.116	.005	.012	-.148	.067	-.016
	5	.087	-.822	.019	-.309	.931	-.442	.188	4.875	.143
Spink	1	.066	-.717	-.237	-.616	.541	1.668	-.303	.029	-.044
	2	.092	-.640	.625	-.426	1.065	2.146	-.058	-.030	2.174
	3	-.043	-.026	.068	.011	.011	-.111	.009	-.033	.109
	4	.051	-.133	.193	.017	.047	-.171	.016	-.073	.090
	5	.158	-.186	-.335	-.182	.227	.576	.106	.276	.191
Averaged overall counties	1	.269	.113	.075	.376	.321	1.314	.241	1.454	.030
	2	.254	-.167	.245	.201	.401	1.527	.133	2.816	.412
	3	-.138	-.020	.086	-.050	.063	-.088	.282	.250	-.011
	4	-.164	-.056	.198	-.003	.116	-.242	.218	.072	-.059
	5	-.095	-.382	.078	.360	.468	1.004	.392	.791	.341

*Cardenas regression estimator.

†Cardenas ratio estimator.

**Maximum likelihood direct proportion estimator.

††Least squares direct proportion estimator.

§Huddleston-Ray estimator.

differences between the larger area, for which the estimator was developed, and the smaller county area, for which estimates are desired.

The ATCOR and XSTAR algorithms transform both the analysis district sample, which is used in training the classifier, and the county data to make them spectrally more alike by correcting for atmospheric and background differences. Not only is a classifier developed that is different from the one that was developed using EDITOR alone, but a different regression estimator is also obtained. The MLEST algorithm, however, provides a transformation that is applied to the training statistics from the analysis district. This transformation changes the classifier to better fit distributions that are present in the county. Thus, the transformed MLEST classifier is used on a particular county, but is inappropriate for the analysis district; hence, the same regression equation is developed on the analysis district by both EDITOR and EDITOR with MLEST. Any differences between EDITOR in county-level estimates and EDITOR with MLEST preprocessing would be caused by the classifier, because the regression equations are the same.

Tables 5-28 through 5-33 present the classification results for the 75 segment samples from the analysis district that were used to train the classifier and develop the regression estimators for EDITOR and for EDITOR with each preprocessor. Also included are similar results for the two county samples, which were both disjoint from the training set.

5.3.1 HOTELLING'S T^2 TEST RESULTS

It has been assumed in this study that there are some fundamental atmospheric and background differences between the six-county analysis district and the individual county to be estimated. Hence, one single overall correction by a preprocessor of the entire six-county area might transform the spectral space in some useful manner (i.e., reduce the effects of haze), but it would not correct for the difference between individual counties and the analysis district. Such differences would still remain, and no real classification improvement was obtained when using XSTAR in this manner. The remaining study used XSTAR and ATCOR in a manner which would have corrected for such

TABLE 5-28.- EDITOR WITHOUT PREPROCESSING

(a) Classification results for 75 segments used to train the classifier

Crop	PCC*
Alfalfa	24.39
Corn	71.97
Wheat	44.44
Oats	47.97
Flax	46.65
Hay cut	12.73
Grass	38.36
Rangeland	69.09
Sunflowers	85.57
Overall PCC	56.41

(b) Classification results for 25 segments from Beadle County and 20 segments from Kingsbury County

Beadle County		Kingsbury County	
Crop	PCC*	Crop	PCC*
Alfalfa	15.23	Alfalfa	31.92
Corn	46.04	Corn	64.20
Wheat	24.42	Wheat	39.84
Oats	16.39	Oats	12.98
Flax		Flax	31.03
Hay cut	3.11	Hay cut	5.56
Grass	5.77	Grass	22.81
Rangeland	59.18	Rangeland	50.28
Sunflowers		Sunflowers	39.47
Overall PCC	36.55	Overall PCC	42.01

*Percentage of correct classification.

TABLE 5-29.- EDITOR WITH XSTAR PREPROCESSING - SINGLE HAZE CORRECTION USED FOR BOTH ANALYSIS DISTRICT SAMPLE AND COUNTY

[Entire 6-county area transformed at once]

(a) Classification results for 75 segments used to train the classifier

Crop	PCC*
Alfalfa	24.51
Corn	71.22
Wheat	37.68
Oats	44.38
Flax	43.41
Hay cut	8.33
Grass	36.26
Rangeland	66.11
Sunflowers	83.56
Overall PCC	53.60

(b) Classification results for 25 segments from Beadle County

Crop	PCC*
Alfalfa	5.26
Corn	46.47
Wheat	22.12
Oats	26.50
Flax	
Hay cut	6.04
Grass	4.81
Rangeland	66.41
Sunflowers	
Overall PCC	39.88

*Percentage of correct classification.

TABLE 5-30.- EDITOR WITH XSTAR PREPROCESSING - ANALYSIS DISTRICT AND COUNTY SEPARATELY CORRECTED FOR HAZE

[County and training areas transformed separately]

(a) Classification results for 75 segments used to train the classifier

Crop	PCC*
Alfalfa	28.55
Corn	59.73
Wheat	41.97
Oats	37.96
Flax	33.87
Hay cut	15.05
Grass	26.59
Rangeland	63.98
Sunflowers	84.73
Overall PCC	50.09

(b) Classification results for 25 segments from Beadle County and 20 segments from Kingsbury County

Beadle County		Kingsbury County	
Crop	PCC*	Crop	PCC*
Alfalfa	11.16	Alfalfa	13.46
Corn	40.60	Corn	56.61
Wheat	20.47	Wheat	41.44
Oats	14.21	Oats	13.74
Flax		Flax	25.29
Hay cut	7.66	Hay cut	0.00
Grass	10.58	Grass	18.83
Rangeland	61.96	Rangeland	43.44
Sunflowers		Sunflowers	40.79
Overall PCC	36.76	Overall PCC	36.74

*Percentage of correct classification.

TABLE 5-31.- EDITOR WITH ATCOR PREPROCESSING

(a) Classification results for 75 segments used to train the classifier

Crop	PCC*
Alfalfa	33.09
Corn	70.13
Wheat	36.20
Oats	11.90
Flax	8.92
Hay cut	14.35
Grass	4.35
Rangeland	56.59
Sunflowers	74.66
Overall PCC	42.90

(b) Classification results for 25 segments from Beadle County and 20 segments from Kingsbury County

Beadle County		Kingsbury County	
Crop	PCC*	Crop	PCC*
Alfalfa	29.78	Alfalfa	18.85
Corn	66.45	Corn	69.63
Wheat	13.66	Wheat	42.08
Oats	2.51	Oats	0.00
Flax		Flax	0.00
Hay cut	17.74	Hay cut	0.00
Grass	0.00	Grass	1.87
Rangeland	49.73	Rangeland	38.45
Sunflowers		Sunflowers	0.00
Overall PCC	38.59	Overall PCC	33.62

*Percentage of correct classification.

TABLE 5-32.- EDITOR WITH MLEST PREPROCESSING

(a) Classification results for 75 segments used to train the classifier†

Crop	PCC*
Alfalfa	24.39
Corn	71.97
Wheat	44.44
Oats	47.97
Flax	46.65
Hay cut	12.73
Grass	38.36
Rangeland	69.09
Sunflowers	85.57
Overall PCC	56.41

†Note that these are exactly the same results as those in table 5-28(a). The classifier is the same.

(b) Classification results for 25 segments from Beadle County and 20 segments from Kingsbury County**

Beadle County		Kingsbury County	
Crop	PCC*	Crop	PCC*
Alfalfa	26.95	Alfalfa	3.46
Corn	52.29	Corn	62.72
Wheat	17.41	Wheat	10.08
Oats	18.77	Oats	12.21
Flax		Flax	5.75
Hay cut	0.16	Hay cut	0.00
Grass	20.43	Grass	32.87
Rangeland	33.71	Rangeland	8.50
Sunflowers		Sunflowers	0.00
Overall PCC	29.17	Overall PCC	28.72

*Percentage of correct classification.
**Note that MLEST has adjusted the classifier for the individual counties, and these results are not the same as those in table 5-28(b).

TABLE 5-33.- EDITOR WITH MLEST PREPROCESSING WITH TRUE PROPORTIONS

(a) Classification results for 75 segments used to train the classifier

Crop	PCC*
Alfalfa	24.39
Corn	71.97
Wheat	44.44
Oats	47.97
Flax	46.65
Hay cut	12.73
Grass	38.36
Rangeland	69.09
Sunflowers	85.57
Overall PCC	56.41

(b) Classification results for 25 segments from Beadle County and 20 segments from Kingsbury County

Beadle County		Kingsbury County	
Crop	PCC*	Crop	PCC*
Alfalfa	50.33	Alfalfa	8.08
Corn	39.59	Corn	63.31
Wheat	9.22	Wheat	14.40
Oats	12.02	Oats	2.29
Flax		Flax	5.17
Hay cut	5.94	Hay cut	0.00
Grass	0.00	Grass	44.68
Rangeland	21.05	Rangeland	2.77
Sunflowers		Sunflowers	0.00
Overall PCC	23.28	Overall PCC	31.46

*Percentage of correct classification.

†The prior probabilities in the MLEST-adjusted classifiers for each county were the actual crop proportions in the counties, rather than the corp proportions in the analysis district.

differences (separately correcting the analysis district and the county), provided that the algorithms were effective in haze correction.

Although the preprocessing methods did not improve classification for all crops for any county, MLEST (without prior knowledge of crop proportions in the county) produced improved or similar classification results in the Beadle County test site for every crop but rangeland. However, MLEST did so poorly in classifying rangeland, which constituted 33 percent of the county sample, that the overall PCC was significantly lower than for EDITOR alone. Similar results were found for XSTAR in the same county when the entire analysis district was corrected at once.

Tables 5-34 and 5-35 show the results of comparing crop-area estimates that are obtained from each preprocessing procedure with those from EDITOR. The Hotelling T^2 test was used in the comparison. Rejecting the null hypothesis would indicate that the preprocessing method produced regression estimates which were significantly different from those produced without preprocessing.

In Beadle County, EDITOR with ATCOR produced regression estimates that were significantly different from those produced by EDITOR alone, but these estimates were better for some crops and worse for others (mixed). EDITOR with XSTAR also produced mixed results in Kingsbury County. Mixed results are undesirable because the preprocessing procedure should produce the same estimates as EDITOR when little or no haze or background difference is present and better estimates for all crops when heavy haze is present. The inconsistency of the XSTAR and ATCOR algorithms indicates that they are either insufficient in their compensation for large differences or that they are unreliable when little difference is present. This conclusion is supported by results published in 1981 by Dave (ref. 8), which suggested that XSTAR corrects mainly for sun angle rather than for haze and that XSTAR contains assumptions which are at variance with findings from the Dave study. The XSTAR and ATCOR algorithms will not be included in the FY 1982 study.

TABLE 5-34.- STRATUM 12 HOTELLING'S T^2 RESULTS OF 25 SEGMENTS IN BEADLE COUNTY

[NULL HYPOTHESIS: EDITOR with preprocessing produces crop area estimates which are the same as those from EDITOR without preprocessing.]

Test	Calculated T^2	Reject/ accept H_0	Results
EDITOR with no preprocessing vs. EDITOR with XSTAR *(I)	12.47	Accept	
EDITOR with no preprocessing vs. EDITOR with XSTAR †(II)	25.31	Accept	
EDITOR with no preprocessing vs. EDITOR with ATCOR †(II)	52.79	Reject	Mixed
EDITOR with no preprocessing vs. EDITOR with MLEST	20.483	Accept	
EDITOR with no preprocessing vs. EDITOR with MLEST (both regression estimators developed on Readle County)	42.79	Accept	
EDITOR with no preprocessing (regression estimator developed on analysis district) vs. EDITOR with MLEST (regression estimator developed on county)	150.97	Reject	Mixed, mostly better
Critical value T^2 (6, 13, .05)	49.23		

TABLE 5-35.- STRATUM 12 HOTELLING'S T^2 RESULTS OF 20 SEGMENTS IN KINGSBURY COUNTY

Test	Calculated T^2	Accept/ reject H_0	Results
EDITOR with no preprocessing vs. EDITOR with XSTAR (II)	111.89	Reject	Mixed
EDITOR with no preprocessing vs. EDITOR with ATCOR	56.24	Accept	
EDITOR with no preprocessing vs. EDITOR with MLEST	26.91	Accept	
Critical value T^2 (8, 14, .05)	86.08		

*I. entire 6-county area transformed at once.
†II. training and county areas transformed separately.

The MLEST procedure consistently produced regression estimates that were not significantly different from those produced by EDITOR alone. These regression estimates were obtained both from the analysis district and from the county. The FY 1980 study showed that a regression developed on a training set was significantly different from a regression developed on an independent set. The only case in which EDITOR with MLEST produced estimates that were better than those produced by EDITOR alone was when estimates from the EDITOR analysis district regression estimator were compared with estimates produced by an estimator developed on the counties where MLEST had transformed the classifier. Because the counties were in effect an independent test set (disjoint from the training segments), such a difference in the regression estimates was probably similar to the training and test estimator differences found in the FY 1980 study and was not due to the use of MLEST.

Below is listed the mean vector used in calculating the Hotelling T^2 statistic when EDITOR with MLEST was used with an estimator that was developed from segments in Beadle County to estimate crop areas for those segments, and EDITOR alone used an estimator developed from sample segments from the whole analysis district.

$$\hat{\mu}_c = \begin{bmatrix} \hat{\mu}_{\text{Range}} \\ \hat{\mu}_{\text{Corn}} \\ \hat{\mu}_{\text{Wheat}} \\ \hat{\mu}_{\text{Oats}} \\ \hat{\mu}_{\text{Hay cut}} \\ \hat{\mu}_{\text{Grass}} \\ \hat{\mu}_{\text{Alfalfa}} \end{bmatrix} = \begin{bmatrix} 13.767 \\ -2.066 \\ 21.923 \\ -4.084 \\ -5.141 \\ 12.350 \\ 18.306 \end{bmatrix}$$

where $\hat{\mu}_{\text{Crop}} = \frac{1}{n} \sum_{i=1}^{n} (|y - \hat{y}_e| - |y - \hat{y}_n|)_i$

In this equation, n is the number of sample segments from Beadle County in stratum 12 (14), $\hat{y}_e$ is the estimate for segment i from EDITOR alone, $\hat{y}_m$ is the estimate for segment i from EDITOR with MLEST, and y is the ground truth for segment i.

When the $\hat{\mu}_{Crop}$ was positive, then MLEST with EDITOR provided estimates which, on the average, were closer to ground truth than EDITOR alone; when the $\hat{\mu}_{Crop}$ was negative, EDITOR provided the better estimates. However, this improvement is probably caused by the different regression equations used, and not by the use of MLEST.

This improvement highlights the fact that in production use with no test segments available, EDITOR with MLEST would use the same regression equations as would EDITOR alone. However, with the use of MLEST, the classifier used to classify the county is not the same as the one used to develop the regression equations. It would be expected that this would affect the estimates made on the county. Any classification improvement caused by the use of MLEST must overcome such degradation in estimator performance.

It should be possible to take advantage of potential improvements if test segments were always available to produce a new estimator or if some other type of estimator were used. Unfortunately, test segments are usually not available, since all available segments are needed to train the classifier adequately.

In order to better evaluate if differences between the crop proportions in the analysis district and the county were affecting the performance of the MLEST classifier, the crop proportions that are listed in tables 5-36 through 5-38 were provided to the MLEST classifier to see if any classification improvement would be obtained.

In Beadle County, two crops are not present at all, and in both counties crop proportions vary widely between analysis district and county for some crops. Classification results using these priors from the county were worse for one county and better for the other when compared to use of MLEST with priors from the training data. Differences in crop proportions between the analysis district and county must then be larger before estimates of the county proportions would improve an MLEST classifier.

TABLE 5-36.- CROP PROPORTIONS OF 75 SEGMENTS IN ANALYSIS DISTRICT

Crop	PPC
Alfalfa	5.41
Corn	13.39
Wheat	9.42
Oats	7.02
Flax	3.27
Hay cut	2.87
Grass	8.23
Rangeland	25.86
Sunflowers	3.95
Other	20.58

TABLE 5-37.- CROP PROPORTIONS OF 25 SEGMENTS IN BEADLE COUNTY

Crop	PPC
Alfalfa	9.0
Corn	16.3
Wheat	4.3
Oats	7.3
Flax	0.0
Hay cut	14.0
Grass	2.1
Rangeland	33.9
Sunflowers	0.0
Other	13.1

TABLE 5-38.- CROP PROPORTIONS OF
20 SEGMENTS IN KINGSBURY COUNTY

Crop	PCC
Alfalfa	6.4
Corn	24.8
Wheat	15.3
Oats	3.2
Flax	4.3
Hay cut	0.4
Grass	20.9
Rangeland	13.3
Sunflowers	1.9
Other	9.5

Although MLEST performed consistently, it never produced estimates that were statistically different from those produced by EDITOR. But there is a question of whether there was actually any difference in the haze level between the analysis district and the two counties. Several methods were used to attempt to answer this question.

Table 5-39 lists the MLEST transformation matrix and vector used to transform the training statistics before classifying each county.

Although the diagonal elements of the A matrices were all close to 1, neither the off-diagonal elements nor the values for the B vector were close to zero. This transformation is not close enough to an identity transformation to say that there is no difference in the distributions of the analysis district and the county; there may be some difference, but not very much.

5.3.2 ATCOR HAZE LEVELS

Table 5-40 displays the haze levels measured by ATCOR for the two acquisitions for the analysis district and county samples.

TABLE 5-39.- MLEST TRANSFORMATION MATRIX A AND VECTOR B FOR BEADLE AND KINGSBURY COUNTIES

(a) Beadle County

Matrix A

1.01	-0.01	-0.14	0.06	0.08	-0.07	0.12	-0.05
0.04	0.89	-0.09	-0.03	0.13	-0.01	0.09	0.00
0.11	-0.14	1.05	-0.07	0.05	-0.08	0.09	-0.03
-0.12	0.01	0.16	0.94	-0.11	-0.09	0.07	-0.03
-0.18	-0.06	-0.06	0.04	1.00	0.16	0.14	-0.09
-0.22	-0.10	-0.07	0.05	-0.08	1.22	0.31	-0.22
-0.22	-0.26	-0.10	-0.02	0.20	0.39	1.05	-0.02
-0.14	-0.26	-0.13	-0.02	0.16	0.42	0.02	1.02

Vector B

0.81	1.21	-1.09	-1.26	0.64	1.39	-0.95	-1.62

(b) Kingsbury County

Matrix A

0.85	0.09	-0.17	0.14	0.09	0.04	-0.04	0.04
-0.09	1.07	-0.16	0.14	0.05	0.16	-0.09	0.03
-0.07	-0.30	0.86	0.07	0.14	-0.11	0.02	0.18
-0.24	-0.29	0.00	0.99	-0.19	0.02	0.12	0.17
-0.19	0.03	-0.16	0.09	1.04	0.21	-0.12	0.14
-0.20	-0.03	-0.09	0.03	0.01	1.34	-0.14	0.15
-0.10	-0.06	-0.18	-0.10	0.15	0.29	1.06	0.04
-0.09	-0.03	-0.10	-0.19	0.06	0.28	0.11	1.04

Vector B

-0.85	-0.72	-1.03	-0.41	-0.20	-0.80	2.07	2.10

TABLE 5-40.- ATCOR-MEASURED HAZE LEVELS

Analysis district.	Haze level	No. of segments
G123 training group:		75
Acquisition 1	0.177	
Acquisition 2	0.236	
Beadle County test group:		25
Acquisition 1	0.250	
Acquisition 2	0.257	
Kingsbury County test group:		20
Acquisition 1	0.113	
Acquisition 2	0.209	

*Haze levels are measured on a scale of 0.000 (no haze) to 1.000 (heavy haze).

Unfortunately, although the haze levels that were measured are useful for comparison, it is not known at what haze level a classifier will first have poor performance caused by haze (table 5-40). Note that there is little difference between the analysis district and either county for acquisition 2, but there is some difference in Beadle County for acquisition 1. XSTAR does not produce a haze diagnostic.

5.3.3 COMPARISON OF REGRESSION LINES

The third method used to look at the presence or absence of haze was to compare regression lines obtained from the training area to those from the Beadle County test area.

Tables 5-41 and 5-42 show the tests for homogeneity of variances and the test for equality of regression lines. The county regression lines were developed on only 14 segments, whereas the analysis district regression lines were developed on 42 segments (both sets from stratum 12 only). As can be seen, homogeneity of variances was rejected for six of seven crops, and the regression lines were not the same for the remaining crop. Any attempt to draw conclusions about haze level from these tests is limited.

The fourth and final attempt to reach an understanding of differences involves an observation of XSTAR results. If classification results from XSTAR with EDITOR (when the whole six-county analysis district was corrected at once) had been worse than those when the analysis district and the county were corrected separately, then it would have been concluded that there was some difference between the county and the analysis district that was not being corrected, although the overall average haze level may have been reduced. In fact, that was not the case. The classification results when the whole area was corrected at once were actually better. The conclusion drawn from this attempt to measure haze and other differences between the county and analysis district is that there is probably some difference present, but it is not a large difference.